Ethnic Geography: Measurement and Evidence*

Roland Hodler† Michele Valsecchi‡ Alberto Vesperoni§

June 3, 2019

Abstract

The effects of ethnic geography, i.e., the distribution of ethnic groups across space, on economic, political and social outcomes are not well understood. We develop a novel index of ethnic segregation that takes both ethnic and spatial distances between individuals into account. Importantly, we can decompose this index into indices of spatial dispersion, generalized ethnic fractionalization, and the alignment of spatial and ethnic distances. We use ethnographic maps, spatially disaggregated population data, and language trees to compute these four indices for around 160 countries. We apply these indices to study the relation between ethnic geography and current economic, political and social outcomes. We document that country level quality of government, income and trust increase with the alignment component of segregation, i.e., with the ratio between the country's actual segregation and the segregation it would have if ethnic groups were represented in each location with population shares identical to their country-level population share. Hence, all else equal, countries where ethnically diverse individuals live farther apart tend to perform better.

Keywords: Ethnic diversity; ethnic geography; segregation; fractionalization; quality of government; economic development.

JEL classification: C43; D63; O10; Z13.

*We acknowledge helpful comments by Magnus Hatlebakk, Mario Jametti, Nadine Ketel, Stelios Michalopoulos, Maria Petrova, Marta Reynal-Querol, Måns Söderbom, Ragnar Torvik, David Yanagizawa-Drott, Ekaterina Zhuravskaya, participants at the 2016 CESifo Workshop on Political Economy, the 2017 ASWEDE conference, the NES CSDSI International Conference "Towards Effective and Equitable Development: the Role of Institutions and Diversity," and seminar participants at IEB Barcelona, CMI Bergen, NHH Bergen, Deakin University, Monash University, Universitat Pompeu Fabra, University of Gothenburg, University of Lugano, University of St.Gallen and University of Zurich. Steve Berggreen-Clausen provided excellent research assistance.

†Department of Economics, University of St.Gallen; CEPR, London; CESifo, Munich; email: roland.hodler@unisg.ch.

‡New Economic School, Moscow; email: mvalsecchi@nes.ru.

§Department of Economics, Alpen-Adria University Klagenfurt; email: alberto.vesperoni@gmail.com.

1 Introduction

There is a vast literature on how a country's ethnic diversity affects economic, political and social outcomes. This literature provides evidence for negative effects of ethnic diversity on, e.g., peace, public goods provision, redistribution, the quality of government, and economic development in general. In these studies, ethnic diversity is typically quantified by indices based on the different ethnic groups' country-wide population shares.[1] By definition, these indices ignore ethnic geography, i.e., the distribution of ethnic groups across space.

Ethnic geography may however play an important role. Consider first a country that is ethnically diverse in all locations. The spatial proximity of ethnically diverse individuals could be a cause of friction and mutual distrust, making cooperation at the local level hard to achieve and possibly leading to dysfunctional communities and local governments.[2] As a result of weak social cohesion and poor governance in most locations, this country might well end up with poor governance and poor economic performance at the national level.

Alternatively, consider a country that is equally ethnically diverse (based on the different ethnic group's country-level population shares), but in which all locations are ethnically homogeneous, as the different ethnic groups are separated from one another. In this country, individual communities may be more functional and local governance better. However, at the country level, divisions may be larger and a sense of community harder to achieve, among other things, because the less cumbersome cooperation and preference aggregation at the local level may make it easier for ethnic groups to recruit resources to fight (peacefully or violently) for their own interests at the national level.

These two hypothetical countries suggest that the effects of ethnic geography on governance at the national level are unclear from a theoretical perspective. The notion that the second (more segregated) country would be worse-off at the national level is consistent with the findings of Alesina and Zhuravskaya (2011), who make an important first step towards taking ethnic geography into account. They construct an index of ethnic segregation that is based on the various ethnic groups' population shares in different subnational units such as regions or provinces. Using this index, which depends on ethnic geography as well as "internal administrative borders which, in turn, are at a government's discretion" (Alesina and Zhuravskaya, 2011, p. 1889), they find that the quality of government is lower in more ethnically segregated countries.

[1] Prominent examples are the index of ethnic fractionalization (e.g., Easterly and Levine 1997, Alesina et al. 2003, Desmet et al. 2012) and the indices of ethnic polarization (e.g., Esteban and Ray 1994, Montalvo and Reynal-Querol 2005). See Alesina and La Ferrara (2005) for a review of the early literature on ethnic diversity and economic performance.

[2] Studies exploiting within-country variation indeed show that higher local ethnic diversity goes hand-in-hand with lower local public goods provision, less trust, less social capital, less cooperation, weaker social norms, and weaker social sanctioning (e.g., Alesina and La Ferrara 2000, 2002, Miguel and Gugerty 2005, Algan et al. 2016, Gershman and Rivera 2017).

We contribute to the literature on ethnic diversity by proposing a set of indices that capture important aspects of ethnic geography. Our first contribution is a methodological one: we derive a new segregation index that is based on both spatial and ethnic distances between pairs of individuals. There is indeed evidence that both these distances matter (see, e.g., White, 1983, for spatial distances and Desmet et al., 2009, for ethnolinguistic distances). To develop our index, we consider a society divided into ethnic or, more generally, social groups and scattered over a territory. The starting point is a general class of indices that are expressions of the relation between a randomly selected pair of individuals. The basic idea is that the relation of two individuals depends on whether they are (i) unlikely to interact personally due to high spatial distance and (ii) unlikely to share a common ethnocultural background due to high ethnic distance. We then uniquely characterize an index from this class via a set of axioms that are intuitive properties of a segregation measure. These axioms capture the notions that segregation is higher when individuals in the same locations are more ethnically homogeneous and when ethnically diverse individuals are located farther apart from one another. Our segregation index can be interpreted as the probability that two randomly selected individuals neither interact personally, nor share a common ethnocultural background.[3]

This index has two prominent features. To understand the first, we make use of the terminology used by Reardon and O'Sullivan (2004). They call segregation measures "a-spatial" if they are based on population shares in administrative units, and "spatial" if they are based on spatial distances between individuals.[4] Our index is a spatial segregation measure. It thereby avoids standard problems of a-spatial segregation measures, in particular the border dependence mentioned by Alesina and Zhuaravskya (2011) and the checkerboard problem (White 1983, Reardon and O'Sullivan 2004).[5] Second, our index can be decomposed into three (sub-)indices: an index of spatial dispersion, a well-known index of generalized ethnic fractionalization (see below), and a measure of the alignment of spatial and ethnic distances between individuals (i.e., ethno-spatial alignment or, simply, alignment hereinafter). Figure 1 illustrates these components and the corresponding properties of our segregation index.

Figure 1 about here

First consider part (a) of this figure, where only the easternmost and the westernmost location are inhabited in the left diagram, and only the two more central locations in the

[3]Such probabilistic interpretation simply requires that ethnic and spatial distances are normalized to take values in the unit interval.

[4]Reardon and Firebaugh (2002) and Reardon and O'Sullivan (2004) review a-spatial and spatial segregation measures, respectively.

[5]There are at least two reasons why overcoming border dependence is important: First, administrative borders are the result of policy choices that may be endogenous to ethnic geography. Second, border-dependent segregation measures can lead to different rankings of ethnic segregation across countries depending on the administrative units used (e.g., provinces/states versus districts). Online Appendix A illustrates border dependence and the checkerboard problem of a-spatial segregation indices.

right diagram. Our index suggests that the society in the right diagram is less segregated than the society in the left diagram because the spatial distance between individuals from ethnically distinct groups (represented by different tones of gray) is lower, all else being equal. This feature is captured by the spatial dispersion component of our segregation index. In part (b) our index suggests that the society in the right diagram is less segregated than the society in the left diagram, because of the lower ethnic distance between individuals in different locations (represented by more similar tones of gray), all else being equal. This is captured by the generalized ethnic fractionalization component. Part (c) illustrates the important role that ethno-spatial alignment plays in our conceptualization. On average, ethnic and spatial distances are identical in the societies in the left and the right diagrams. However, in the society in the left diagram ethno-spatial alignment is high, as individuals that are ethnically most distant are also located furthest apart compared to the benchmark where each ethnicity is proportionally represented in each location. Ethno-spatial alignment is lower in the society in the right diagram, where ethnically distant individuals live spatially closer to one another while spatially distant individuals are ethnically closer.

Our second contribution is applied in nature. We compute these four indices of ethnic geography for 161 countries from all over the world.[6] We define as ethnic groups the language groups listed in the Ethnologue (Gordon, 2005). To measure ethnolinguistic distances, we rely on the Ethnologue's language trees. To measure spatial distances, we use the World Language Mapping System's (WLMS) map that represents the traditional homeland of each language group listed in the Ethnologue.[7] To measure local population, we use population density maps from CIESIN (2019). In addition, we construct a simple map based on global land cover data that should proxy for the exogenous component of the spatial distribution of a country's population.

We use our indices in cross-country regressions to improve our understanding of the role that ethnic geography plays in economic, political and social outcomes around the globe. Our indices are well suited to this purpose thanks to the various precautions we took in designing and computing them. First, they are based on spatial distances rather than administrative borders. They are therefore not driven by the drawing of administrative borders, which is a policy choice that may be endogenous to ethnic geography. Second, we have computed these indices for many countries, so that we have a sample with almost full global coverage. Third, the reliance on a map of the traditional homelands of language groups goes some way in ensuring that our the indices are not driven by recent (voluntary

[6] We do not compute our indices for small countries with a current population of less than 250,000 or a land surface area of less than 5,000 km^2.

[7] The WMLS aims at representing "the region within each country, which is the traditional homeland of each indigenous language" (WMLS, version 19, n.p.). There is one caveat. In some former colonies where many Europeans settled, native groups got largely displaced and the WLMS map shows the new territories of these language groups as their traditional homelands. We discuss and address this caveat in Sections 3.3 and 3.4, where we exclude 26 former settler colonies in some robustness tests.

or forced) migration and urbanization, in particular when combining this map with the soil-suitability-based proxy for the spatial distribution of the population.

We first focus on the associations between our index of ethnic segregation on the one hand, and the quality of government, incomes and generalized trust on the other hand. We find a negative relation between ethnic segregation and the quality of government, similar to Alesina and Zhuravskaya (2011) with their index of a-spatial segregation in their sample of 97 countries. We further find that our index of ethnic segregation tends to be negatively associated with incomes too, but unrelated to generalized trust.

More importantly, we study the relation between the three components of our index of ethnic segregation – ethnic fractionalization, spatial dispersion and ethno-spatial alignment – and these outcome variables. Ethnic fractionalization tends to be associated with lower-quality government and lower incomes. This association is not overly robust when controlling for some biological, climatic and geographical variables that may shape ethnic diversity and ethnic geography. It is nonetheless the main reason for the negative relation between our index of ethnic segregation on the one hand, and the quality of government and incomes on the other hand. Spatial dispersion is basically unrelated to all of our outcome variables. Most strikingly, we find a positive and statistically significant association between the alignment of ethnic and spatial distances between individuals, and the quality of government, incomes and trust. Hence, conditionally on ethnic fractionalization and spatial dispersion, societies in which ethnically diverse people live far apart are, on average, better governed, richer and more trusting. Such conditionality is what differentiates the analysis of segregation from the analysis of alignment: higher alignment does not indicate higher segregation (relative to another country), but a level of segregation higher than what the country would have had, had all ethnic groups been represented in all locations with population shares identical to their country shares. Unlike in Figure 1(c), such "benchmark" segregation does vary across countries and we thus control for it.

Our work is related to other contributions on the measurement of segregation that incorporate the spatial dimension. Several contributions introduce spatial distances into well-known a-spatial models of segregation (e.g., Jakubs 1981 for the dissimilarity index; White 1983 for the isolation index; or Reardon and O'Sullivan 2004 for the dissimilarity index, the Theil index and the interaction index). Moreover, Echenique and Fryer Jr (2007) develop a segregation index based on proximity in networks.[8] To our knowledge, there is, however, no other segregation measure that presents both ethnic/social and spatial distances in the same framework.[9]

[8]In their model spatial distances are binary, but the degree of isolation of an individual depends on the isolation of every other individual in the network. Blumenstock and Fratamico (2013) also rely on network data for providing a-spatial segregation measures.

[9]Methodologically, our approach is in the tradition of exposure measurement, being loosely based on the isolation-interaction models of Bell (1954), White (1983), and Philipson (1993). Most axiomatic work on segregation focuses on another class of models, known as evenness indices (e.g., Hutchens 2004, Chakravarty and Silber 2007, and Frankel and Volij 2011). While some evenness measures are extended

Our framework is also related to prominent models of fractionalization and polarization (e.g., Esteban and Ray 1994, Duclos et al. 2004, Bossert et al. 2011), as we introduce ethnic/social distances in the very same way they do. In particular, the generalized ethnic fractionalization component of our ethnic segregation index coincides with the generalized fractionalization index introduced by Greenberg (1956) and later axiomatized by Bossert et al. (2011), which in turn is equivalent to the standard fractionalization index when ethnic distances are binary.[10]

As mentioned earlier, our paper is related to the extensive literature on the relation between ethnic diversity and economic, political and social outcomes. We contribute to this literature by developing, computing and applying our spatial index of ethnic segregation and its three sub-indices – all with global coverage. There are two complementary strands of the literature that also rely on ethnographic maps to study the role of ethnic geography. The first of these strands chooses subnational ethnographic regions as units of analysis. Prominent examples include studies on the relation between the location of ethnic groups and conflict (e.g., Cederman et al. 2009, Weidmann 2009, Michalopoulos and Papaioannou 2016, König et al. 2017), on the effect of pre-colonial and current institutions on development (Michalopoulos and Papaioannou 2013, 2014), and on ethnic favoritism (De Luca et al. 2018). These contributions provide interesting insights into the effect of ethnic geography on within-country variation while our segregation index allows for comparing ethnic geography across countries and understanding the country-level effects of ethnic geography.

Just as we do, contributions to the second strand combine ethnographic maps with population density maps to construct country-level measures of ethnic diversity and ethnic geography. Matuszeki and Schneider (2006) compute a measure of average subnational ethnic fractionalization, and study how this measure relates to conflict at the country level. Desmet et al. (2016) develop a measure that captures the average exposure of an individual to members of the country's different ethnic groups with an emphasis on weighting this exposure according to the representation of these groups at the individual's location. They study how this measure relates to public goods provision. There are two main differences between these approaches and ours: First, we focus on conceptualizing ethnic segregation and introducing the novel concept of ethno-spatial alignment, while they extend the fractionalization framework. Matuszeki and Schneider (2006) do so in a straightforward way, and Desmet et al. (2016) by introducing population weights in a non-linear fashion. Second, spatial (and ethnic) distances are continuous in our approach,

to introduce spatial distances, they do not lend themselves naturally to the introduction of both spatial and ethnic distances.

[10] From a purely mathematical view point, the generalized fractionalization index axiomatized in Bossert et al. (2011) is an unnormalized Gini index. Analogously, our segregation index can be seen as a particular type of multivariate Gini index (see, e.g., Gajdos and Weymark 2005). However, as it violates standard majorization criteria of multivariate inequality measurement, it should not be interpreted as an inequality measure.

but binary in Matuszeki and Schneider (2006) and Desmet et al. (2016). We thus see our spatial segregation index as complementary to their measures, which capture alternative important aspects of ethnic diversity and ethnic geography.[11]

Section 2 presents the theoretical framework, derives our segregation index, and establishes its decomposability into indices of generalized ethnic fractionalization, spatial dispersion, and ethno-spatial alignment. Section 3 first explains the data and the methodology used to construct our four indices of ethnic geography. It then offers a first look at these indices and presents our cross-country evidence. Section 4 concludes.

2 Development of indices of ethnic geography

2.1 General model

A population is partitioned into n ethnic or, more generally, social groups $G := \{1, \ldots, n\}$ and distributed over t locations on a territory $T := \{1, \ldots, t\}$. We generally assume $t \geq n \geq 3$ so that (i) there is significant ethnic heterogeneity; (ii) there are at least as many locations as groups, so that it is possible that no individuals of different groups share the same location.

Denote by $\mu_p^g \in [0,1]$ the share of population that corresponds to group $g \in G$ in location $p \in T$. Let $\mu_p := \sum_{g \in G} \mu_p^g$ and $\mu^g := \sum_{p \in T} \mu_p^g$ be the total population shares of location $p \in T$ and group $g \in G$ respectively, where $\sum_{p \in T} \mu_p = \sum_{g \in G} \mu^g = 1$. Then, the $n \times t$ matrix of population shares

$$\mu := \begin{bmatrix} \mu_1^1 & \cdots & \mu_t^1 \\ \vdots & \ddots & \vdots \\ \mu_1^n & \cdots & \mu_t^n \end{bmatrix}$$

defines a mass distribution, and the space of all mass distributions $\mathcal{M}$ is the subset of $[0,1]^{t \times n}$ such that the restrictions above are satisfied. For any pair of locations $p, q \in T$, let $\lambda_{p,q} \in [0,1]$ be the spatial distance between them, where we generally assume $\lambda_{p,q} = 0$ if $p = q$ and $\lambda_{p,q} = \lambda_{q,p}$. A spatial distribution is defined by the $t \times t$ matrix of spatial

[11] Montalvo and Reynal-Querol (2016) use ethnographic maps to look at ethnic geography by computing ethnic fractionalization in grid cells of different sizes. Alesina et al. (2016) and Guariso and Rogall (2016) use ethnographic maps to measure inequality across ethnic groups and to study the country-level effects of between-group inequality on economic development and conflict, respectively. Due to the focus of these studies, they take neither the spatial distances between individuals from different ethnic homelands nor the linguistic distances between individuals from different ethnic groups into account.

distances between all pairs of locations

$$\lambda := \begin{bmatrix} \lambda_{1,1} & \cdots & \lambda_{1,t} \\ \vdots & \ddots & \vdots \\ \lambda_{t,1} & \cdots & \lambda_{t,t} \end{bmatrix},$$

and the space of all spatial distributions $\mathcal{L}$ is the subset of $[0,1]^{t\times t}$ such that the restrictions above are satisfied. For any pair of groups $g, h \in G$, let $\gamma^{g,h} \in [0,1]$ be the ethnic distance between them, where we generally assume $\gamma^{g,h} = 0$ if $g = h$ and $\gamma^{g,h} = \gamma^{h,g}$. The $n \times n$ matrix of ethnic distances between all pairs of groups

$$\gamma := \begin{bmatrix} \gamma^{1,1} & \cdots & \gamma^{1,n} \\ \vdots & \ddots & \vdots \\ \gamma^{n,1} & \cdots & \gamma^{n,n} \end{bmatrix}$$

defines an ethnic distribution, and the space of all ethnic distributions $\mathcal{G}$ is the subset of $[0,1]^{n\times n}$ such that the restrictions above are satisfied. Finally, a joint distribution is a triple of mass, spatial and ethnic distributions, and an index is a function $S : ([0,1]^{t\times n}, [0,1]^{t\times t}, [0,1]^{n\times n}) \to \mathbb{R}_+$, where $S(\mu, \lambda, \gamma)$ quantifies some property of the joint distribution $(\mu, \lambda, \gamma) \in (\mathcal{M}, \mathcal{L}, \mathcal{G})$.

To give meaning to our framework we now impose some more structure. We assume (a relevant feature of) the relation between each pair of individuals is determined by the distances between their groups and locations.[12] For each pair of individuals that inhabit locations $p, q \in T$ and belong to groups $g, h \in G$, we quantify the relation between them by $\pi(\lambda_{p,q}, \gamma^{g,h})$, where the function $\pi : [0,1]^2 \to \mathbb{R}_+$ is continuous and non-decreasing in each argument and satisfies $\pi(0,0) = 0$. Among the various interpretations of the function π, one possibility is to see it as the degree of alienation (i.e., lack of common interests) between a pair of individuals, which naturally increases with their spatial and ethnic distances. Given this, we consider the class of indices that are expression of the relation between a randomly selected pair of individuals, taking the form

$$S(\mu, \lambda, \gamma) = \sum_{(p,q)\in T^2} \sum_{(g,h)\in G^2} \mu_p^g \mu_q^h \pi(\lambda_{p,q}, \gamma^{g,h}) \tag{1}$$

for each joint distribution $(\mu, \lambda, \gamma) \in (\mathcal{M}, \mathcal{L}, \mathcal{G})$ and any function π that satisfies the above restrictions.

In the next section we will introduce a set of axioms that pin down a particular index (up to positive scalar multiplication) from class (1) as our segregation index. This coincides with identifying a specific function π that is suitable for the measurement of

[12] For related approaches, see Esteban and Ray (1994), Duclos et al. (2004), and Bossert et al. (2011).

segregation. As our initial restrictions on π are weak (π can be, e.g., logarithmic, exponential, multiplicative, additive, etc.), class (1) is vast. Nevertheless, the focus on class (1) considerably narrows the set of indices under consideration by taking pairs of individuals as the relevant unit of analysis and by imposing that any pair's contribution to segregation depends on their spatial and ethnic distances only.[13] We are not concerned by these restrictions for three reasons intimately related to the interpretability/decomposability of the index and the crucial issues of border dependence and data availability in applications:

1. We think of segregation as a measure of the extent to which ethnically diverse *individuals* are located far apart, which captures the notion that society becomes more segregated when the interaction between ethnically diverse individuals becomes less likely. This allows for an intuitive interpretation of our segregation measure in terms of probabilistic interaction and a meaningful decomposition into well-known concepts such as ethnic fractionalization and spatial dispersion (see Section 2.3).

2. We deliberately take spatial (and ethnic) *distances* as primitives of the model in order to build a segregation measure that is independent of borders between locations (and ethnic groups) that are arbitrarily drawn to achieve a discrete categorization. The focus on distances rather than individual attributes is in line with our empirical application where ethnic distances are not derived as differences between cardinal attributes of ethnic groups but rather as a function of the number of nodes that their languages share on a linguistic tree (see Section 3.1).

3. As our unit of analysis is the pair of individuals and we want to focus on their ethnic/spatial distances, function π could only be generalized by making it dependent on some elements of the mass distribution μ. However, by introducing some element of μ in function π, we would implicitly assume that the relation between two individuals is discontinuous at some borders between locations (or ethnic groups) and any generalization of π would therefore (re-)introduce border dependence "through the back door."[14]

[13] To see this, one can rewrite form (1) as a function of distances between pairs of individuals rather than between pairs of groups and locations. With some abuse of notation, let $\lambda_{i,j}$ and $\gamma^{i,j}$ denote the spatial and ethnic distances between each pair of individuals i, j from a finite population P. Then, $S = (1/|P|^2) \sum_{(i,j)\in P^2} \pi(\lambda_{i,j}, \gamma^{i,j})$.

[14] As pointed out in Footnote 13, class (1) can be written as a function of spatial and ethnic distances between pairs of individuals. In applications, categorizing individuals in a limited number of locations and ethnicities (i.e., introducing arbitrary borders) is a necessary approximation. Ideally, this should not lead to systematic biases in the computation of the index. While these biases are minimal for class (1) as they tend to "average out" due to the linearity in each element of μ, they would be magnified if we had some element of μ in function π due to the non-linearity.

2.2 Axiomatization of the segregation index

We now introduce a set of axioms that are desirable properties of a segregation measure. For simplicity of exposition, these desirable properties are defined through simple examples of distributions with two or three mass points. The first two axioms consider pairs of groups and locations, thereby focusing on obtaining ethnic homogeneity within a location. In particular, segregation should increase when the population becomes ethnically homogeneous in all locations, so that there is no local interaction between ethnically diverse individuals. Axiom 1 formalizes this property and, in addition, requires it to hold when the ethnic distance between groups is reduced by an arbitrarily small amount.

Axiom 1 (Local ethnic homogeneity and ethnic distances) *Data:* Consider a joint distribution $(\mu, \lambda, \gamma) \in (\mathcal{M}, \mathcal{L}, \mathcal{G})$ with two locations $p, q \in T$ and two groups $g, h \in G$ such that

$$\mu_p^g = \mu_p^h = \mu_q^h = 1/3 \text{ with } \lambda_{p,q} > 0 \text{ and } \gamma^{g,h} > 0,$$

and let $\tilde{\mu} \in \mathcal{M}$, $\tilde{\gamma} \in \mathcal{G}$ and $\epsilon > 0$ satisfy

$$\tilde{\mu}_p^g = \mu_p^g \text{ and } \tilde{\mu}_q^h = \mu_p^h + \mu_q^h \text{ with } \tilde{\gamma}^{g,h} = \gamma^{g,h} - \epsilon.$$

Statement: We require $S(\mu, \lambda, \gamma) < S(\tilde{\mu}, \lambda, \tilde{\gamma})$ for ϵ arbitrarily small.

Let us discuss Axiom 1, whose distributions are depicted in Figure 2(a). There are two locations (left and right) and two ethnic groups (represented by dark and light tones of gray). Initially, in distribution (μ, λ, γ), two-thirds of the population are in the left location, whose ethnic composition is perfectly balanced (half dark, half light), while the remaining one-third of the population is in the right location and is homogeneously dark. Given this, we transfer all individuals of the dark group into the right location, so that the left location becomes homogeneously light while the right location remains homogeneously dark. Moreover, we reduce the ethnic distance between the light and the dark group by an arbitrarily small amount ϵ (represented by the slightly lighter tone of gray of the dark group in the right diagram). Axiom 1 requires segregation to increase as a consequence of this transformation. Intuitively, the axiom considers a trade off between ethnic homogeneity within locations and the ethnic distance across groups, requiring the former to dominate the trade off when the reduction in ethnic distance is arbitrarily small.

Figure 2 about here

Axiom 2 is very similar to Axiom 1. It is based on the same initial distribution and the same transfer of population from the left to the right location. The only difference is that, instead of reducing the ethnic distance between the light and the dark groups, we

reduce the spatial distance between the left and right locations by an arbitrarily small amount.

Axiom 2 (Local ethnic homogeneity and spatial distances) *Data:* Consider a joint distribution $(\mu, \lambda, \gamma) \in (\mathcal{M}, \mathcal{L}, \mathcal{G})$ with two locations $p, q \in T$ and two groups $g, h \in G$ such that

$$\mu_p^g = \mu_p^h = \mu_q^h = 1/3 \text{ with } \lambda_{p,q} > 0 \text{ and } \gamma^{g,h} > 0.$$

and let $\tilde{\mu} \in \mathcal{M}$, $\tilde{\lambda} \in \mathcal{L}$ and $\epsilon > 0$ satisfy

$$\tilde{\mu}_p^g = \mu_p^g \text{ and } \tilde{\mu}_q^h = \mu_p^h + \mu_q^h \text{ with } \tilde{\lambda}_{p,q} = \lambda_{p,q} - \epsilon.$$

Statement: We require $S(\mu, \lambda, \gamma) < S(\tilde{\mu}, \tilde{\lambda}, \gamma)$ for ϵ arbitrarily small.

These distributions are depicted in Figure 2(b). Intuitively, this axiom considers a trade off between ethnic homogeneity within locations and the spatial distance across locations, requiring the former to dominate the trade off when the reduction in the spatial distance is arbitrarily small.

The next two axioms are still inspired by the generally desirable property that segregation should increase whenever the interaction between ethnically diverse individuals becomes less likely. However, unlike Axioms 1 and 2, they consider triples of groups and locations, thereby focusing on changes in distributions that foster the alignment of spatial and ethnic distances across pairs of individuals. The basic idea is that, to obtain higher segregation, closely located pairs of individuals should be ethnically closer, while ethnically distant pairs should be spatially further apart. Axioms 3 and 4 formalize this idea.

Axiom 3 (Alignment of ethnic distances) *Data:* Consider any joint distribution $(\mu, \lambda, \gamma) \in (\mathcal{M}, \mathcal{L}, \mathcal{G})$ with three locations $p, q, r \in T$ and three groups $g, h, i \in G$ such that

$$\begin{aligned} \mu_p^g = \mu_q^h = \mu_r^i &= 1/3, \\ \lambda_{p,q} > \lambda_{q,r} > 0 \text{ and } \lambda_{p,r} &= \lambda_{p,q} + \lambda_{q,r}, \\ \gamma^{g,h} = \gamma^{h,i} &= \gamma^{g,i}/2 > 0, \end{aligned}$$

and let $\tilde{\gamma} \in \mathcal{G}$ and $\epsilon > 0$ satisfy

$$\tilde{\gamma}^{g,i} = \gamma^{g,i},\ \tilde{\gamma}^{g,h} = \gamma^{g,h} + \epsilon,\ \tilde{\gamma}^{h,i} = \gamma^{h,i} - \epsilon.$$

Statement: We require $S(\mu, \lambda, \gamma) < S(\mu, \lambda, \tilde{\gamma})$ for all $\epsilon \in (0, \gamma^{h,i})$.

Let us discuss Axiom 3, whose distributions are depicted in Figure 2(c). The population mass is uniformly distributed on three locations (left, central and right) and three

ethnic groups (represented by dark, medium and light tones of gray), where the left location is homogeneously light, the central location is homogeneously medium and the right location is homogeneously dark. The three locations are on a line, where the central location is closer to the right than to the left. Regarding ethnic distances, the medium group is halfway between the other two groups in the left diagram representing distribution (μ, λ, γ). Axiom 3 requires segregation to increase when we change ethnic distances so that the medium group becomes ethnically closer to the dark group (represented by the darker tone of gray of the middle location in the right diagram). This is intuitive: as the medium group already inhabits a location that is spatially closer to the location of the dark group than to the location of the light group, the interaction between ethnically diverse individuals becomes less likely.

Axiom 4 (Alignment of spatial distances) *Data:* Consider any joint distribution $(\mu, \lambda, \gamma) \in (\mathcal{M}, \mathcal{L}, \mathcal{G})$ with three locations $p, q, r \in T$ and three groups $g, h, i \in G$ such that

$$\begin{aligned} \mu_p^g = \mu_q^h = \mu_r^i &= 1/3, \\ \lambda_{p,q} = \lambda_{q,r} &= \lambda_{p,r}/2 > 0, \\ \gamma^{g,h} > \gamma^{h,i} > 0 \text{ and } &\gamma^{g,i} = \gamma^{g,h} + \gamma^{h,i}, \end{aligned}$$

and let $\tilde{\lambda} \in \mathcal{L}$ and $\epsilon > 0$ satisfy

$$\tilde{\lambda}_{p,r} = \lambda_{p,r},\ \tilde{\lambda}_{p,q} = \lambda_{p,q} + \epsilon,\ \tilde{\lambda}_{q,r} = \lambda_{q,r} - \epsilon.$$

Statement: We require $S(\mu, \lambda, \gamma) < S(\mu, \tilde{\lambda}, \gamma)$ for all $\epsilon \in (0, \lambda_{q,r})$.

Figure 2(d) represents Axiom 4 graphically. Again, there are three locations respectively inhabited by three equally sized ethnic groups. The medium group is ethnically closer to the dark group than to the light, while the central location is halfway between the right and the left location. Axiom 4 requires segregation to increase if the central location is moved closer to the right location. Similarly to the previous axiom, the intuition is that as the spatial distance between ethnically diverse individuals increases, their interaction becomes less likely.

Our four axioms identify our segregation index from the class of measures (1):[15]

Theorem 1 *An index from class (1) satisfies Axioms 1-4 if and only if it takes the form*

$$S(\mu, \lambda, \gamma) = \sum_{(p,q) \in T^2} \sum_{(g,h) \in G^2} \mu_p^g \mu_q^h \lambda_{p,q} \gamma^{g,h}, \tag{2}$$

up to a positive scalar multiplication.

[15]The proof of Theorem is 1 in the Appendix.

This theorem implies that our segregation index always provides unambiguous rankings of joint distributions $(\mu, \lambda, \gamma) \in (\mathcal{M}, \mathcal{L}, \mathcal{G})$. Further, it implies that ethnic and spatial distances are complementary forces in the determination of the relation of a pair of individuals, so that segregation is high only if pairs of individuals that are ethnically heterogeneous are systematically located apart from each other.

For any $\lambda_{p,q} \in [0,1]$ and $\gamma^{g,h} \in [0,1]$, the function $\pi(\lambda_{p,q}, \gamma^{g,h}) = \lambda_{p,q}\gamma^{g,h}$ always takes value in $[0,1]$. It can thus be interpreted probabilistically. Intuitively, the relation between two individuals depends on (i) whether they do not interact personally and (ii) whether they do not share a common ethnocultural background. Given this, it is natural to interpret the function π as the probability that *both* these events are realized, where the spatial distance $\lambda_{p,q}$ is the probability of event (i) and the ethnic distance $\gamma^{g,h}$ is the probability of event (ii). Then, our segregation index S represents the probability that two randomly selected individuals neither interact personally nor share an ethnocultural background.

2.3 Decomposition of the segregation index

By construction, our segregation index is strongly related to the fractionalization literature. Letting $\mathbf{1}_t$ be the spatial distribution where the spatial distance between each pair of locations is equal to 1 (so that space "does not matter"),[16] it is easy to show that our index is equivalent to the generalized fractionalization index by Bossert et al. (2011),

$$F(\mu, \gamma) := S(\mu, \mathbf{1}_t, \gamma) = \sum_{(g,h)\in G^2} \mu^g \mu^h \gamma^{g,h}. \tag{3}$$

This generalized fractionalization index represents the average ethnic distance between pairs of individuals, and can be interpreted as the probability that two randomly selected individuals do not share a common ethnocultural background. If we also impose ethnic distances to take value in $\{0,1\}$, our index reduces to the standard fractionalization index, which has been widely applied to measure ethnic fractionalization based on categorical data (see, e.g., Alesina et al. 2004 and references therein).[17]

Applying the same reasoning to the other dimension, and letting $\mathbf{1}_n$ be the ethnic distribution where the distance between each pair of groups is 1 (so that ethnicity "does not matter"),[18] we can define the spatial dispersion index as

$$D(\mu, \lambda) := S(\mu, \lambda, \mathbf{1}_n) = \sum_{(p,q)\in T^2} \mu_p \mu_q \lambda_{p,q}. \tag{4}$$

[16] As the spatial distance between a location and itself is also equal to 1, it follows that $\mathbf{1}_t \notin \mathcal{L}$.

[17] To see this, let $\mathbf{1}_n^0 \in \mathcal{G}$ be the ethnic distribution where $\gamma^{g,h} = 1$ if $h \neq g$ and $\gamma^{g,g} = 0$ for each $g \in G$, so that $F(\mu, \mathbf{1}_n^0) = S(\mu, \mathbf{1}_t, \mathbf{1}_n^0) = 1 - \sum_{g\in G} (\mu^g)^2$. This is the standard fractionalization index, which indicates the probability that two randomly selected individuals belong to different ethnic groups.

[18] As the ethnic the distance between a group and itself is also equal to 1, it follows that $\mathbf{1}_n \notin \mathcal{G}$.

This index measures the average spatial distance between pairs of individuals and can be interpreted as the probability that two randomly selected individuals will not interact personally.

Our segregation index tends to be high if spatial distances between locations and ethnic distances between groups are high, i.e., when F and D are high. Moreover, it also depends on the alignment between spatial and ethnic distances, i.e., on whether a high spatial distance between two individuals tends to go hand-in-hand with a high ethnic distance between them. For each $\mu \in \mathcal{M}$, denote by $\overline{\mu} \in \mathcal{M}$ the benchmark mass distribution corresponding to μ, where (i) groups and locations have the same mass as in μ, i.e., $\overline{\mu}^g = \mu^g$ and $\overline{\mu}_p = \mu_p$ for all $g \in G$ and $p \in T$; and (ii) groups are proportionally represented at each location, i.e., $\overline{\mu}^g_p / \overline{\mu}_p = \overline{\mu}^g$ for all $g \in G$ and $p \in T$. Accordingly, we refer to $S(\overline{\mu}, \lambda, \gamma)$ as the benchmark segregation of $S(\mu, \lambda, \gamma)$, and we propose as a measure of ethno-spatial alignment

$$A(\mu, \lambda, \gamma) := \begin{cases} S(\mu, \lambda, \gamma)/S(\overline{\mu}, \lambda, \gamma) & \text{if } S(\overline{\mu}, \lambda, \gamma) > 0, \\ 1 & \text{if } S(\overline{\mu}, \lambda, \gamma) = 0. \end{cases} \tag{5}$$

Given our probabilistic interpretation of S, A can be seen as a likelihood ratio: it is the probability that two randomly selected individuals do not interact personally and do not share an ethnocultural background given mass distribution μ, relative to the probability of the same event given the corresponding benchmark mass distribution $\overline{\mu}$, which is identical to μ except that ethnic groups are represented at each location proportionally to their aggregate shares. Intuitively, focusing on the likelihood ratio should "neutralize" the magnitude effects of average spatial and ethnic distances. In fact, $A(\mu, k\lambda, k'\gamma) = A(\mu, \lambda, \gamma)$ for all $k, k' > 0$, while $S(\mu, k\lambda, k'\gamma) = kk' S(\mu, \lambda, \gamma)$ for all $k, k' > 0$. Hence, our measure of alignment satisfies scale invariance with respect to both spatial and ethnic distances, while our segregation index does not. Other properties of our measure of alignment directly follow from the axioms in the previous section, which are all satisfied in the sense that alignment increases whenever segregation increases.

Lastly, we show how the various measures are related to one other:[19]

Proposition 1 *It holds that*

$$S(\mu, \lambda, \gamma) = \begin{cases} F(\mu, \gamma) D(\mu, \lambda) A(\mu, \lambda, \gamma) & \textit{if } F(\mu, \gamma) > 0 \textit{ and } D(\mu, \lambda) > 0, \\ 0 & \textit{if } F(\mu, \gamma) = 0 \textit{ or } D(\mu, \lambda) = 0. \end{cases} \tag{6}$$

This proposition shows that our segregation index S can be decomposed into the generalized ethnic fractionalization index F, the spatial dispersion index D, and the alignment

[19] The proof of Proposition 1 is in the Appendix.

index A in a multiplicative fashion.[20] Moreover, Proposition 1 and equation (5) imply

$$S(\overline{\mu}, \lambda, \gamma) = F(\mu, \gamma)D(\mu, \lambda).$$

That is, benchmark segregation is equal to the product of generalized ethnic fractionalization and spatial dispersion.

3 Application

3.1 Data and computation of our indices of ethnic geography

We aim at computing our indices of ethnic geography, i.e., the segregation index and its three components, for a large and diverse set of countries from all over the world. For these countries, we need information on locations and ethnic groups, so that we can then derive mass distribution μ, spatial distribution λ, and ethnic distribution γ. These distributions are the inputs required for the computation of our indices.

We therefore combine two data sources. First, we use the Ethnologue (Gordon, 2005), which provides a comprehensive list of the world's known living languages. We consider the language groups listed in the Ethnologue as ethnic groups. It is important to remember that language is more than just a communication device. Common language often implies common ancestry, homeland, cultural heritage, norms, and values.[21] The advantages in relying on the Ethnologue for classifying ethnic groups are fourfold: First, the Ethnologue provides a comprehensive rather than a selective list of ethnolinguistic groups. Second, the Ethnologue provides linguistic trees for the different language families which show the historical relation between all languages. These linguistic trees are thus helpful in measuring linguistic distances between ethnic groups. Third, the World Language Mapping System (WLMS, version 19) provides a map representing the homelands of the language groups in the Ethnologue. This map allows measuring spatial distances between locations inhabited by different groups. Last, but not least, this map represents "the region within each country, which is the traditional homeland of each indigenous language" (WLMS, version 19, n.p.), while populations living away from their traditional homelands, e.g., migrations to cities and refugees, are not mapped.[22]

The second data source is the population density map from the Gridded Population

[20] We discuss in Online Appendix B how this decomposition relates to the interpretation of our segregation index as a geometric projection and to a decomposition of S based on the Euclidean norms of vectors of spatial and ethnic distances.

[21] Desmet et al. (2017) find that ethnic identity is an important determinant of responses to many questions on cultural norms, values and preferences in the World Value Surveys.

[22] There is, however, one caveat. In some former colonies where many Europeans settled, native groups were largely displaced and the WLMS map shows the new territories of these language groups as their traditional homelands. We discuss and address this caveat in Sections 3.3 and 3.4, where we exclude 26 former settler colonies in some robustness tests.

of the World (GPW, version 4), which is based on population census tables and administrative boundaries, and provided by CIESIN (2016). For some robustness tests, we use an alternative map that should proxy for the exogenous spatial distribution of a country's population shaped just by biological, climatic and geographical forces. This map is based on geo-coded information on global land cover from the U.S. Geological Survey (2000) and the assumption that a country's population is equally distributed across its entire area potentially suitable for agriculture.[23]

We take as ethnic groups in each country all the language groups with more than 100 native speakers listed in the Ethnologue and with a homeland mapped within this country. The median and average number of ethnic groups per country are 9 and 41, respectively. There is however a lot of variability in the number of groups: Some countries (15 out of 161 in our sample) have only one ethnic group, while Papua New Guinea, Indonesia and Nigeria have 767, 640 and 450 ethnic groups, respectively.

To determine locations, we use grid cells of 0.5×0.5 arc minutes (corresponding to around 9×9 km near the equator) and cut them at country borders and at the boundaries between different ethnic homelands. We thereby get "proper" squared cells as well as smaller "squiggly" cells (due to country borders or ethnic homeland boundaries). We take each of these (proper or squiggly) cells as a location.

To determine the mass distribution μ, we rely on the population density map from GPW (or the land-cover-based alternative described above). Let m, m_p and m_p^g denote the total population of a country, the population in cell p and the population of language group g in cell p, respectively. For cells p that are part of a homeland of a single language group g, it is straightforward that $m_p^g = m_p$, where m_p is given by GPW. The WMLS map suggests that most homelands have only one language group, but other homelands contain more than one and up to seven language groups. We find that 90 percent of our proper and squiggly cells belong to the homeland of a single group. The remaining 10 percent of our cells belong to ethnic homelands of multiple ethnic groups. Let n_p denote the number of ethnic groups whose ethnic homeland includes cell p. We find that for 9 percent of cells $n_p = 2$, while $n_p > 2$ for 1 percent of cells. For these groups and cells, we simply assume $m_p^g = \frac{m_p}{n_p}$, where m_p is given by GPW and n_p by WMLS.[24] We then compute population shares as $\mu_p^g = \frac{m_p^g}{m}$, where $m = \sum_{p \in T} m_p$.

Figure 3 illustrates the ethnic homelands and the grid cells for Togo (left) and Benin (right). Moreover, it indicates the population in each proper and squiggly cell. We will

[23]The map used is the "Global Land Cover Characteristics Data Base Version 2.0," which classifies land in 18 categories. We classify all areas as potentially suitable for agriculture except deserts, semi-deserts, glaciers, and tundra.

[24]This simple rule may lead us to overestimate the local population of very small language groups, which is the main reason for dropping languages spoken by no more than 100 individuals.

come back to this figure soon.

Add Figure 3 around here

To derive the spatial distribution λ, we use ArcGIS to determine the centroid of each (proper or squiggly) cell p. We then use the latitude and the longitude of these centroids to compute the geodesic distance between any two cells p and q of any given country. We normalize these geodesic distances by the average geodesic distance across all cell-pairs of this country. We use the resulting relative spatial distances $\lambda_{p,q}$ in the computation of our indices.[25]

To derive the ethnic distribution γ, we rely on the Ethnologue's linguistic trees for the different language families. Linguistic trees characterize each language by a series of nodes and thereby contain information about the evolution of languages and the historical relation between ethnolinguistic groups. Two languages share no common node if they belong to different language families, e.g., the Indo-European and the Uralic language family. Such coarse divisions suggest that the language groups separated early and interacted little. In contrast, languages with many common nodes, e.g., Norwegian and Swedish, suggest that the language groups separated late or interacted regularly. Following Fearon (2003), it has become common practice to calculate linguistic distance between groups as a function of the number of common nodes of their languages and to use the linguistic distance between groups as a proxy for their cultural distance more broadly defined. We follow Putterman and Weil (2010, Appendix C) in defining the ethnic distance between ethnic groups g and h as

$$\gamma^{g,h} := 1 - \sqrt{2\tilde{\eta}^{g,h}/(\eta^g + \eta^h)},$$

where η^i is the number of nodes of language $i \in \{g, h\}$ and $\tilde{\eta}^{g,h}$ the number of common nodes.[26]

Using mass distribution μ, spatial distribution λ, and ethnic distribution γ, we derive our indices of ethnic geography for 161 countries with a land surface area of more than 5,000 km^2 and a current population of more than 250,000.[27]

[25] In the absence of this normalization, the indices of ethnic segregation and spatial dispersion would tend to increase in a country's area. Notice, however, that this normalization does not affect our indices of generalized ethnic fractionalization and ethno-spatial alignment. We show in Online Appendix G that our results, in particular the positive effects of ethno-spatial alignment, are robust to using absolute spatial distances in the computation of our indices.

[26] Fearon (2003) proposes a slightly different formula. We show in Online Appendix G that our results are robust to using this formula.

[27] See Online Appendix C for a list of the 161 countries for which we provide our indices of ethnic geography. Besides small countries, we also exclude Austria, because the homelands in the WMLS map cover only a small portion of the area, and Serbia, because of the many changes to its borders.

3.2 A first look at our indices

Table 1 provides some summary statistics for our indices of ethnic geography, and Figure 4 provides scatter plots illustrating the empirical relation between our index of ethnic segregation and its three components.

Add Table 1 and Figure 4 around here

The twelve most ethnically segregated countries according to our index of ethnic segregation are (in decreasing order of segregation) Papua New Guinea, Nigeria, Cameroon, India, Laos, Belize, Namibia, Equatorial Guinea, East Timor, Chad, Bolivia and Liberia. The two scatter plots in the top row of Figure 4 show positive correlations between ethnic segregation, on the one hand, and spatial dispersion or ethnic fractionalization, on the other hand. They suggest that Cameroon (CMR) and, in particular, Equatorial Guinea (GNQ) are among the most ethnically segregated countries mainly because they are highly spatially dispersed, while Papua New Guinea (PNG), Chad (TCD) and Bolivia (BOL) are among the most ethnically segregated countries mainly because they are highly ethnically fractionalized.

These two scatter plots also illustrate that neither high ethnic fractionalization, nor high spatial dispersion is sufficient for high ethnic segregation. Good examples are Spain (ESP) and Suriname (SUR): Spain has high spatial dispersion, as its major metropolitan areas are quite far from one another. It has also low generalized ethnic fractionalization, as most people speak Spanish or a closely related language (like Catalan or Galician, but unlike Basque). As a result, its ethnic segregation is relatively low despite the high spatial dispersion. Suriname is a country with high linguistic distances between various ethnic groups and, therefore, high generalized ethnic fractionalization. But it has also low spatial dispersion with most people living in the capital city or nearby, such that ethnic segregation is relatively low nevertheless.

The scatter plot on the bottom left of Figure 4 shows the relation between our index of ethnic segregation and the alignment between ethnic and spatial distances. It documents an empirically negative relation between ethnic segregation and ethno-spatial alignment. We have seen in Proposition 1 in Section 2 that, all else being equal, segregation increases with ethno-spatial alignment. This scatter plot now shows that, all else not being equal, more aligned countries tend to be less ethnically segregated. The scatter plot on the bottom right of Figure 4 shows that, as we would expect, the relation between ethnic segregation and ethno-spatial alignment becomes positive once we partial out benchmark segregation (which corresponds to $F \times D$).

Norway is the country with the highest ethno-spatial alignment. Most people speak Norwegian, which is a language from the Indo-European language family, and live relatively close to one another in the South of the country (e.g., around Bergen or Oslo).

There are however some small language groups that speak Kven Finnish and Sami. Like Finnish, these languages belong to the Uralic language family. Moreover, the homelands of these language groups are in the far North of Norway. These people are therefore both linguistically and spatially very far from the Norwegian speakers in the South, such that the linguistic distance of a pair of individuals is a very good predictor of the spatial distance, and vice versa.

Interestingly, there are also countries where alignment is less than one, implying that the ethnic distance between spatially distant pairs of individuals tends to be smaller than the ethnic distance between spatially close pairs of individuals. One example is Turkmenistan, where the Turkmen are the largest language group. Moreover, there are three minority groups, speaking Balochi, Kurdish, and Uzbek. Balochi and Kurdish belong to the Indo-European language family, while Turkmen and Uzbek belong to the Altaic language family. Because the homelands of the two Indo-European languages are in fairly central and densely populated areas, pairs of linguistically diverse individuals live on average closer to one another than pairs of individuals speaking the same or very similar languages.

Of course, Norway and Turkmenistan differ in many dimensions. Let us therefore look at Benin and Togo, which differ in their ethno-spatial alignment, but are similar along many other dimensions. They are neighboring countries located in West Africa, with comparable climatic, geographic and demographic characteristics. Moreover, they were both French colonies after WWI, became independent in 1960, and started their post-colonial history in tumultuous ways that culminated in coups by French-trained military figures: Mathieu Kérékou in Benin and Gnassingbé Eyedéma in Togo (Meredith, 2005). These autocrats both managed to stay in power for many years. Benin and Togo are also comparable in terms of generalized ethnic fractionalization (between the median and the third quartile of our sample) and spatial dispersion (above the third quartile). Ethno-spatial alignment is however considerably higher in Benin (1.35, which is above the third quartile) than in Togo (1.11, which is below the median). Figure 3 shows the different ethnic homelands and the main language groups to which these ethnic homelands belong to. Ethno-spatial alignment is relatively high in Benin as there is a relatively clear divide between Kwa speaking groups in the south, Defoid speaking groups in the center, Gur speaking groups in the north, and some smaller groups speaking very different languages in the north east. As a result of this divide, linguistically distant individuals tended to live far apart from one another. In contrast, ethno-spatial alignment is relatively low in Togo, mainly because there are Gur and Kwa speaking groups in the country's south, its center and its north. As a result of these large and widespread language groups, linguistically distant individuals often lived relatively close to one another.

Finally, in Figure 5, we compare our spatial index of ethnic segregation to the a-spatial

index of ethnic segregation by Alesina and Zhuravskaya (2011).

Add Figure 5 around here

The correlation between the two indices is relatively high (0.621), but far from perfect.[28] This latter finding is not surprising given the conceptual differences between the two indices as well as the reliance on different data in the computation of the indices. The use of data from different sources is by itself a consequence of the conceptual differences: our spatial index requires data on ethnic and spatial distances, while a-spatial indices are based on the population shares of different ethnolinguistic groups in different subnational units.

3.3 Empirical approach

In a next step, we use our indices of ethnic geography to see whether they are helpful in predicting and understanding differences in the quality of government and economic outcomes across countries. The use of cross-country regressions is common in the literature on the effects of ethnic heterogeneity, as is the caveat that the estimated coefficients may not necessarily represent causal effects despite efforts to reduce the risks of omitted variable bias and reverse causality.

In most specifications, we use continent fixed effects, thereby restricting our attention to variation within continents. To further address omitted variable bias, we control for variables that are known determinants of ethnic heterogeneity or ethnic geography, and may have direct effects on current economic and institutional outcomes. In particular, we focus on four sets of additional control variables, related to the countries' climate, terrain and history. First, we add temperature, precipitation and absolute latitude to control for climate. Nettle (1998) argues that the duration of the growing season is a key determinant of the number of ethnic groups in a territory, and calculates this duration based on temperature and precipitation. In addition, climate is known to have more direct effects on economic outcomes as well (e.g., Dell et al., 2012).

Second, we control for terrain ruggedness and its interaction with a dummy variable for Africa as well the mean and standard deviation of elevation and soil suitability for agriculture. Nunn and Puga (2012) argue that rugged terrain has generally negative effects on economic development, but positive effects in Africa, where rugged terrain offered protection against slave raiders. Nunn (2008) further argues that the slave trade promoted ethnic and political fragmentation and had negative effects on economic development. Michalopoulos (2012) shows that geographic variability, which he proxies by the mean

[28] Online Appendix D (Table D.1) reports correlation coefficients between our four indices and various alternative indices of ethnic diversity. Notice the low correlation between ethno-spatial alignment and all other indices.

and standard deviation of elevation and soil suitability, is a key determinant of ethnic diversity across and within countries. At the same time, land productivity is likely to have direct economic effects as well.

Third, turning to historical variables, we control for the time elapsed since the agricultural transition, the migratory distance to Addis Ababa (Ethiopia), and its squared term, and dummy variables indicating the former colonizer (if any). Ahlerup and Olsson (2012) argue that the agricultural transition had strong effects on population density and ethnic heterogeneity; and the biological and geographical factors that led to the early emergence of sedentary agriculture may well have shaped economic development. Migratory distance from the cradle of humankind in East Africa is a predictor for the duration of human settlement. Ahlerup and Olsson (2012) argue that ethnic diversity increases with this duration. In addition, Ashraf and Galor (2013) show that genetic diversity is a decreasing function of the migratory distance from East Africa, and that economic development is a hump-shaped function of genetic diversity.

Fourth, we control for dummy variables indicating the former European colonizer (if any), as there is considerable evidence that the random drawing of borders and divide-and-rule strategies by the colonial powers shaped ethnic heterogeneity and ethnic geography, and had long-term effects on economic and political outcomes (e.g., Alesina et al., 2011, Michalopoulos and Papaioannou, 2016).[29]

Our choice to base the indices on a map of traditional homelands should reduce (but not eliminate) the risk of reverse causality. We further tackle two related concerns. The first is that our indices of ethnic geography are based on current population density data. We therefore present additional results for indices based on the assumption that the (historical) population of any given country was uniformly distributed across all areas that are potentially suitable for agriculture.

The second concern is that, for some settler colonies, the WLMS map does not indicate the traditional homeland of some displaced native language groups, but their new territory. We therefore present specifications in which we exclude 26 settler colonies, defined as countries where more than 10 percent of the year 2000 population have ancestors from former European colonial powers according to Putterman and Weil's (2010) world migration matrix.

[29]See Online Appendix E for more information about the control variables. We take many of the control variables from Ashraf and Galor (2013). Following them and many others, we exclude from our sample the relatively young countries Montenegro and South Sudan as well as Palestine and Taiwan, which are not UN member states, leaving us with a sample of 157 countries with a land surface area of more than 5,000 km^2 and a current population of more than 250,000.

3.4 Cross-country evidence

3.4.1 Ethnic geography and the rule of law

Inspired by Alesina and Zhuravskaya (2011), we first look at the rule of law as a measure of the quality of government. This measure is provided by the World Bank Governance Indicators. By construction, it has a mean of 0 and a standard deviation of 1. In our sample, its 2010 value has a mean of -0.203 and a standard deviation of 0.988. Table 2 shows our results. The columns differ in the set of control variables used. The top panel presents estimates using our index of ethnic segregation. The panel in the middle disaggregates this index into ethno-spatial alignment and benchmark segregation, and the bottom panel replaces it with all three components: ethno-spatial alignment, generalized ethnic fractionalization, and spatial dispersion.

Table 2 around here

We see in column (1) that the rule of law is negatively associated with our index of ethnic segregation in the absence of control variables. This negative association is consistent with the findings by Alesina and Zhuravskaya (2011). When disaggregating this index into ethno-spatial alignment and benchmark segregation, we find a negative relation between benchmark segregation and the rule of law. It is this negative relation that drives the negative relation between our segregation index and the rule of law. In contrast, ethno-spatial alignment is positively associated with the rule of law. This latter result is novel, as is the concept of ethno-spatial alignment itself. Hence, given the level of benchmark segregation, a country has a better rule of law if individuals from very different groups live far apart from one another. Importantly, conditioning on benchmark segregation is akin to say that a country has a better rule of law if individuals from very different groups live far from one another, relative to where they would have lived, had all groups been represented in each location with population shares equal to their country ones.

When decomposing our segregation index into its three components, the coefficient estimates on ethno-spatial alignment remain almost unchanged. Further, we find – consistent with the previous literature (e.g., Alesina et al., 2003) – that the rule of law is negatively associated with fractionalization. The high correlation between fractionalization and segregation (see Figure 4) implies that this negative association between fractionalization and the rule of law is the main reason for the negative association between (benchmark) segregation and the rule of law. In contrast, we find no statistically significant association between spatial dispersion and the rule of law.

In column (2), we add continent fixed effects. The associations of the rule of law with segregation (in the top panel) and fractionalization (in the bottom panel) remain

negative, but their magnitude drops by around 50 percent. In contrast, the association with alignment changes little (in the middle and the bottom panel). The point estimates suggest that an increase of alignment by one standard deviation is associated with an increase in the rule of law by 20–22 percent of a standard deviation.

In columns (3)–(6), we add the four sets of additional control variables discussed above. We see that the association between fractionalization and the rule of law varies quite strongly with the set of control variables, while the association between ethno-spatial alignment and the rule of law is relatively stable in magnitude and remains statistically significant in all these specifications.[30]

In column (7), we use our indices computed based on the assumption of equal population across habitable areas. The coefficient estimate on ethno-spatial alignment becomes even slightly higher, but – not surprisingly – less precisely estimated.

In column (8), we exclude the 26 former colonies where more than 10 percent of the current population has ancestors from former European colonial powers according to Putterman and Weil's (2010) world migration matrix.[31] The coefficient estimate on ethno-spatial alignment drops slightly, but remains broadly similar as in the full sample. Hence, our results are not driven by former colonies where many Europeans settled and where native groups may have been displaced. We conclude that high alignment between ethnic and spatial distances goes hand-in-hand with a high quality of government.

3.4.2 Ethnic geography and income

We now look at the association between ethnic geography and income, measured by the log of expenditure-side real GDP per capita in USD in 2010 from the Penn World Tables 9.0. Table 3, which shows the results, is organized in the same way as the previous table.

Table 3 around here

The results are similar as well. Ethnic segregation is negatively associated with income, but this association is imprecisely estimated in many specifications. We find a similar pattern for generalized ethnic fractionalization when we decompose segregation into its three components. Moreover, the association between spatial dispersion and income is not statistically significant (with the exception of column (7)). The association between ethno-spatial alignment and income is however positive and statistically significant in all specifications. The point estimates in the middle and the bottom panel of column (2) suggest that an increase in alignment by one standard deviation is associated with an

[30] When adding all 24 control variables jointly, the coefficient estimate on ethno-spatial alignment becomes 0.35 (with a p-value of 0.055) in the bottom-panel specification.

[31] These 26 former colonies are 20 Latin American countries, "Neo-Europe" (i.e., Australia, Canada, New Zealand and the United States) plus Namibia and South Africa. In Online Appendix G, we present additional robustness tests in which we exclude each continent individually or just "Neo-Europe."

increase in income by 21–24 percent.

Hence, high alignment between ethnic and spatial distances goes hand-in-hand with high quality of government as well as high incomes today. This pattern also holds true when comparing Benin and Togo. Remember that these neighboring countries are similar along many dimensions, but ethno-spatial alignment is higher in Benin. Our data show that Benin indeed does better in terms of quality of government (-0.70 vs -0.91) and income per capita (USD 1,728 vs USD 1,214).[32]

3.4.3 Ethnic geography and trust

These strong associations raise the question about possible mechanisms linking traditional ethno-spatial alignment with current quality of government and current incomes. The within-country studies by Alesina and La Ferrara (2000, 2002), Miguel and Gugerty (2005), and Algan et al. (2016) document that high local ethnic diversity leads to or is at least associated with low social capital and lack of trust. High ethno-spatial alignment implies that ethnic diversity tends to be low in most locations (conditional on the level of ethnic fractionalization). As a result, trust may be higher in countries with high ethno-spatial alignment.

We use generalized trust from the World Values Surveys in the 1981–2008 time period (taken from Ashraf and Galor, 2013) to look at the role of trust. Generalized trust is measured as the fraction of people answering "most people can be trusted" (as opposed to "can't be too careful") when asked the standard trust question (see Online Appendix E for details). We have coverage for 77 countries, which implies a drop in sample size by around 50 percent. Table 4 presents the associations between our indices of historical ethnic geography and trust.

Table 4 around here

Ethno-spatial alignment is indeed positively associated with generalized trust in all specifications. The point estimates in the middle and the bottom panel of column (2) suggest that an increase in alignment by one standard deviation is associated with an increase in trust by 34–40 percent of a standard deviation. In contrast, ethnic segregation, benchmark segregation, ethnic fractionalization and spatial dispersion are all unrelated to trust (with changes in signs across columns and high p-values throughout).

In Online Appendix F (Table F.1), we provide further evidence for the possibility that trust could be a mechanism linking ethno-spatial alignment to a high quality of government and high incomes. There, we show that the associations between ethno-spatial alignment, on the one hand, and the quality of government and income, on the other hand, become considerably weaker once we control for trust. These findings are consistent with the notion that more aligned countries might be performing better because of higher trust.

[32]The data on trust, introduced in Section 4.3, is missing for Benin and Togo.

3.4.4 Robustness

We present various robustness tests in Online Appendix G. Tables G.1–G.3 show that our results are robust to the exclusion of individual continents, "Neo-Europe" (i.e., Australia, Canada, New Zealand and the United States), all ethnically homogenous countries, or outliers. Table G.4 shows that our results are robust to the use of alternative measures of governance and income. Table G.5 shows that our results are robust to computing our indices based on alternative measures of ethnic and spatial distances, e.g., ethnic distances computed based on Fearon's (2003) formula or absolute spatial distances. Table G.6 shows that our results for ethno-spatial alignment are not an artifact of non-linear effects of ethnic fractionalization and spatial dispersion. Table G.7 shows that the use of weighted least squares leads to very similar results, implying that our main results are not just driven by small countries. Table G.8 presents results using poisson pseudo-maximum likelihood (PPML). This robustness test is of particular interest as ethnic segregation is equal to the product of its three components (see Proposition 1). It is reassuring that the PPML estimates suggest the same general pattern, in particular sizeable positive effects of alignment on the quality of government, income and trust.

Furthermore, in Online Appendix H (Tables H.1–H.3), we report various specifications that include alternative indices of ethnic diversity or ethnic geography as additional right-hand side variables. The associations of ethno-spatial alignment with the rule of law, income and trust remain positive and statistically significant in all specifications.

4 Conclusions

To better understand the role of ethnic geography and to mitigate well-known problems of a-spatial segregation measures, we have developed a new segregation index that is based on ethnic distances between groups and spatial distances between locations rather than categorical data on ethnic groups and administrative units. The decomposition of our segregation index reveals that it corresponds to the product of generalized ethnic fractionalization, spatial dispersion, and the alignment between ethnic and spatial distances. This ethno-spatial alignment is a novel concept that captures, broadly speaking, whether ethnically different individuals tend to live far from each other, relative to the situation where all groups appeared in each location with population shares equal to their country ones.

We have computed these indices using linguistic trees as well as maps of traditional ethnic homelands and spatially disaggregated population data. Using these indices in cross-country regressions suggests, among other things, that countries with higher ethno-spatial alignment tend to be better governed, richer, and more trusting.

We expect our indices to become useful in future work on the role of ethnic geography

in shaping economic, political and social outcomes across countries. However, we also hope to speak to the rapidly growing literature that uses ethnic homelands (or grid cells) as units of analysis to achieve convincing identification strategies. To this literature, we would like to convey the message that local economic, political or social outcomes in any given ethnic homeland may well depend on the broader ethnic geography of the area or country in which this homeland is located.

Of course, the indices we have developed can also be applied to measure the ethnic geography of cities. For example, one could use our segregation index instead of a-spatial measures to compare segregation across US metropolitan areas or within metropolitan areas over time. Given that our indices allow for non-categorical ethnicity data, they may be even more attractive in studying the ethnic geography of emerging African mega-cities, where there is typically great variability in ethnic distances across pairs of individuals.

Finally, we would like to stress that our theoretical framework is not specific to the ethnic dimension. Instead of categorizing individuals by ethnic groups and measuring linguistic distances, future research could focus on other social or socio-economic cleavages that are believed to be salient in a particular setting.

Appendix: Proofs

Proof of Theorem 1: It is easy to verify that our segregation index (2) belongs to class (1) and satisfies Axioms 1-4. Let us show that, if an index belongs to class (1) and satisfies Axioms 1-4, then it must take the form (2) up to a positive scalar multiplication. Take any index from class (1) and let $a, b > 0$ be any scalars, where a is spatial distance and b is ethnic distance in what follows. By Axiom 1, for $\epsilon > 0$ arbitrarily small,

$$\pi(a, b) + \pi(0, b) + \pi(a, 0) < 2\pi(a, b - \epsilon).$$

Letting $a \to 0$, by continuity of π and $\pi(0, 0) = 0$, we obtain at the limit

$$\pi(0, b) \leq \pi(0, b - \epsilon).$$

Then, since π is non-decreasing, $\pi(0, b)$ must be constant in b; and by $\pi(0, 0) = 0$ we must have

$$\pi(0, b) = 0 \text{ for all } b \geq 0. \tag{7}$$

Similarly, by Axiom 2, for $\epsilon > 0$ arbitrarily small,

$$\pi(a, b) + \pi(0, b) + \pi(a, 0) < 2\pi(a - \epsilon, b),$$

so that letting $b \to 0$ by the same arguments we obtain

$$\pi(a, 0) = 0 \text{ for all } a \geq 0. \tag{8}$$

Keeping our interpretation of a as spatial distance and b as ethnic distance, let $c > 0$ be any scalar that represents another spatial distance in the following. By Axiom 3, for all $\epsilon \in (0, b)$

$$\pi(a, b) + \pi(c, b) < \pi(a, b + \epsilon) + \pi(c, b - \epsilon) \text{ if } c < a,$$
$$\pi(a, b) + \pi(c, b) > \pi(a, b + \epsilon) + \pi(c, b - \epsilon) \text{ if } c > a,$$

hence by continuity of π

$$\pi(a, b) + \pi(c, b) = \pi(a, b + \epsilon) + \pi(c, b - \epsilon) \text{ if } c = a.$$

Rearranging terms this leads to

$$\pi(a, b) = \frac{\pi(a, b + \epsilon) + \pi(a, b - \epsilon)}{2} \text{ for all } \epsilon \in (0, b),$$

hence π must be linear in the second argument. Jointly with (7) and (8), this implies $\pi(a,b) = \phi(a)b$ for all $a, b \geq 0$, where $\phi : [0,1] \to \mathbb{R}_+$ is some continuous non-decreasing function that satisfies $\phi(0) = 0$. Similarly, by Axiom 4 (interpreting a as spatial distance, b as ethnic distance and c as another ethnic distance), for all $\epsilon \in (0, b)$

$$\pi(b,a) + \pi(b,c) = \pi(b+\epsilon, a) + \pi(b-\epsilon, c) \text{ if } c = a,$$

hence π must also be linear in the first argument. It follows that $\phi(a) = ka$ for some $k > 0$, and we obtain $\pi(a,b) = kab$ for all $a, b \geq 0$. □

Proof of Proposition 1: It is straightforward that, if $F(\mu,\gamma) = 0$ or $D(\mu,\lambda) = 0$, we must have $S(\mu,\lambda,\gamma) = 0$. To see this, note that $F(\mu,\gamma) = 0$ implies $\gamma^{g,h} = 0$ for all $g, h \in G$ with $\mu^g, \mu^h > 0$. Similarly, $D(\mu,\lambda) = 0$ implies $\lambda_{p,q} = 0$ for all $p, q \in T$ with $\mu_p, \mu_q > 0$. Then, if $F(\mu,\gamma) = 0$ or $D(\mu,\lambda) = 0$, there is either zero spatial distance or zero ethnic distance between each pair of individuals, which implies $S(\mu,\lambda,\gamma) = 0$ by the multiplicative form of π.

We now show that, if $F(\mu,\gamma) > 0$ and $D(\mu,\lambda) > 0$, we must have

$$S(\mu,\lambda,\gamma) = F(\mu,\gamma)D(\mu,\lambda)A(\mu,\lambda,\gamma).$$

By the definition of $A(\mu,\lambda,\gamma)$, this is true if and only if

$$S(\overline{\mu},\lambda,\gamma) = F(\mu,\gamma)D(\mu,\lambda), \tag{9}$$

where the uniform mass distribution $\overline{\mu}$ corresponding to μ is such that (i) $\overline{\mu}^g = \mu^g$ and $\overline{\mu}_p = \mu_p$ for all $g \in G$ and $p \in T$; and (ii) $\overline{\mu}^g_p / \overline{\mu}_p = \overline{\mu}^g$ for all $g \in G$ and $p \in T$. Combining the definition of our index with (ii) we obtain

$$\begin{aligned} S(\overline{\mu},\lambda,\gamma) &= \sum_{(p,q)\in T^2} \sum_{(g,h)\in G^2} \left(\overline{\mu}_p \overline{\mu}^g\right)\left(\overline{\mu}_q \overline{\mu}^h\right) \lambda_{p,q}\gamma^{g,h} \\ &= \left(\sum_{(p,q)\in T^2} \overline{\mu}_p \overline{\mu}_q \lambda_{p,q}\right)\left(\sum_{(g,h)\in G^2} \overline{\mu}^g \overline{\mu}^h \gamma^{g,h}\right), \end{aligned}$$

which together with (i) implies (9). □

References

Ahlerup, Pelle, and Ola Olsson, "The Roots of Ethnic Diversity," *Journal of Economic Growth*, 17 (2012), 71–102.

Alesina, Alberto, Arnaud Devleeschauwer, William Easterly, Sergio Kurlat, and Romain Wacziarg, "Fractionalization," *Journal of Economic Growth*, 8 (2003), 155–194.

Alesina, Alberto, William Easterly, and Janina Matuszeski, "Artificial States," *Journal of the European Economic Association*, 9 (2011), 246–277.

Alesina, Alberto, and Eliana La Ferrara, "Participation in Heterogeneous Communities," *Quarterly Journal of Economics*, 115 (2000), 847–904.

Alesina, Alberto, and Eliana La Ferrara, "Who Trusts Others?" *Journal of Public Economics*, 85 (2002), 207–234.

Alesina, Alberto, and Eliana La Ferrara, "Ethnic Diversity and Economic Performance," *Journal of Economic Literature*, 43 (2005), 762–800.

Alesina, Alberto, Stelios Michalopoulos, and Elias Papaioannou, "Ethnic Inequality," *Journal of Political Economy*, 124 (2016), 428–488.

Alesina, Alberto, and Ekaterina Zhuravskaya, "Segregation and the Quality of Government in a Cross Section of Countries," *American Economic Review*, 101 (2011), 1872–1911.

Algan, Yann, Camille Hémet, and David Laitin, "The Social Effects of Ethnic Diversity at the Local Level: A Natural Experiment with Exogenous Residential Allocation," *Journal of Political Economy*, 124 (2016), 696–733.

Ashraf, Quamrul, and Oded Galor, "The 'Out of Africa' Hypothesis, Human Genetic Diversity, and Comparative Economic Development," *American Economic Review*, 103 (2013), 1–46.

Bell, Wendell, "A Probability Model for the Measurement of Ecological Segregation," *Social Forces*, 32 (1954), 357–364.

Blumenstock, Joshua, and Lauren Fratamico, "Social and Spatial Ethnic Segregation: A Framework for Analyzing Segregation with Large-Scale Spatial Network Data," *Proceedings of the 4th Annual Symposium on Computing for Development*, 4 (2013), 11.

Bossert, Walter, Conchita D'Ambrosio, and Eliana La Ferrara, "A Generalized Index of Fractionalization," *Economica*, 78 (2011), 723–750.

Cederman, Lars-Erik, Halvard Buhaug, and Jan K. Rød, "Ethno-Nationalist Dyads and Civil War: A GIS-based Analysis," *Journal of Conflict Resolution*, 53 (2009), 496–525.

Center for International Earth Science Information Network at Columbia University (CIESIN), *Gridded Population of the World, Version 4*, Palisades, NY (2016).

Chakravarty, Satya R., and Jacques Silber, "A Generalized Index of Employment Segregation," *Mathematical Social Sciences*, 53 (2007), 185–195.

De Luca, Giacomo, Roland Hodler, Paul A. Raschky, and Michele Valsecchi, "Ethnic

Favoritism: An Axiom of Politics?" *Journal of Development Economics*, 132 (2018), 115–129.

Dell, Melissa, Benjamin F. Jones, and Benjamin A. Olken, "Temperature Shocks and Economic Growth: Evidence from the Last Half Century," *American Economic Journal: Macroeconomics*, 4 (2012), 66–95.

Desmet, Klaus, Joseph Gomes, and Ignacio Ortuño-Ortín, "The Geography of Linguistic Diversity and the Provision of Public Goods," CEPR Discussion Paper 11683 (2016).

Desmet, Klaus, Ignacio Ortuño-Ortín, and Romain Wacziarg, "The Political Economy of Linguistic Cleavages," *Journal of Development Economics*, 97 (2012), 322–338.

Desmet, Klaus, Ignacio Ortuño-Ortín, and Romain Wacziarg, "Culture, Ethnicity and Diversity," *American Economic Review*, 107 (2017), 2479–2513.

Desmet, Klaus, Shlomo Weber, and Ignacio Ortuño-Ortín, "Linguistic Diversity and Redistribution," *Journal of the European Economic Association*, 7 (2009), 1291–1318.

Duclos, Jean-Yves, Joan Esteban, and Debraj Ray, "Polarization: Concepts, Measurement, Estimation," *Econometrica*, 72 (2004), 1737–1772.

Easterly, William, and Ross Levine, "Africa's Growth Tragedy: Policies and Ethnic Divisions," *Quarterly Journal of Economics*, 112 (1997), 1203–1250.

Echenique, Federico, and Roland G. Fryer, Jr., "A Measure of Segregation Based on Social Interactions," *Quarterly Journal of Economics*, 122 (2007), 441–485.

Esteban, Joan, Laura Mayoral, and Debraj Ray, "Ethnicity and Conflict: An Empirical Study," *American Economic Review*, 102 (2012), 1310–1342.

Esteban, Joan, and Debraj Ray, "On the Measurement of Polarization," *Econometrica*, 62 (1994), 819–851.

Fearon, James D., "Ethnic and Cultural Diversity by Country," *Journal of Economic Growth*, 8 (2003), 195–222.

Frankel, David M., and Oscar Volij, "Measuring School Segregation," *Journal of Economic Theory*, 146 (2011), 1–38.

Gajdos, Thibault, and John A. Weymark, "Multidimensional generalized Gini indices," *Economic Theory*, 26 (2005), 471–496.

Gershman, Boris, and Diego Rivera, "Subnational Diversity in Sub-Saharan Africa: Insights from a New Dataset," Mimeo (2017).

Gordon, Raymond G., Jr., *Ethnologue: Languages of the World* (Dallas: SIL International, 2005).

Greenberg, Joseph H., "The Measurement of Linguistic Diversity," *Language*, 32 (1956), 109–115.

Guariso, Andrea, and Thorsten Rogall, "Rainfall Inequality, Political Power, and Ethnic Conflict in Africa," Mimeo (2016)

Hutchens, Robert M., "One Measure of Segregation," *International Economic Review*, 45 (2004), 555–578.

Jakubs, John F., "A Distance-Based Segregation Index," *Socio-Economic Planning Sciences*, 15 (1981), 129–136.

König, Michael D., Dominic Rohner, Mathias Thoenig, and Fabrizio Zilibotti, "Networks in Conflict: Theory and Evidence from the Great War of Africa." *Econometrica*, 85 (2017), 1093–1132.

Matuszeki, Janina, and Frank Schneider, "Patterns of Ethnic Group Segregation and Civil Conflict," Mimeo (2006).

Meredith, Martin, *The Fate of Africa: A History of the Continent Since Independence* (New York: Free Press, 2005).

Michalopoulos, Stelios, "The Origins of Ethnolinguistic Diversity," *American Economic Review*, 102 (2012), 1508–1539.

Michalopoulos, Stelios, and Elias Papaioannou, "Pre-Colonial Ethnic Institutions and Contemporary African Development," *Econometrica*, 81 (2013), 113–152.

Michalopoulos, Stelios, and Elias Papaioannou, "National Institutions and Subnational Development in Africa," *Quarterly Journal of Economics*, 129 (2014), 151–213.

Michalopoulos, Stelios, and Elias Papaioannou, "The Long-Run Effects of the Scramble for Africa," *American Economic Review*, 106 (2016), 1802–1848.

Miguel, Edward, and Mary Kay Gugerty, "Ethnic Diversity, Social Sanctions, and Public Goods in Kenya," *Journal of Public Economics*, 89 (2005), 2325–2368.

Montalvo, Jose G., and Marta Reynal-Querol, "Ethnic Polarization, Potential Conflict, and Civil Wars," *American Economic Review*, 95 (2005), 796–816.

Montalvo, Jose G., and Marta Reynal-Querol, "Ethnic Diversity and Growth: Revisiting the Evidence," Mimeo (2016).

Nettle, Daniel, "Explaining Global Patterns of Language Diversity," *Journal of Anthropological Archaeology*, 17 (1998), 354–374.

Nunn, Nathan, "The Long-term Effects of Africa's Slave Trades," *Quarterly Journal of Economics*, 123 (2008), 139–176.

Nunn, Nathan, and Diego Puga, "Ruggedness: The Blessing of Bad Geography in Africa," *Review of Economics and Statistics*, 94 (2012), 20–36.

Philipson, Tomas, "Social Welfare and Measurement of Segregation," *Journal of Economic Theory*, 60 (1993), 322–334.

Putterman, Louis, and David N. Weil, "Post-1500 Population Flows and The Long-Run Determinants of Economic Growth and Inequality," *Quarterly Journal of Economics*, 125 (2010), 1627–1682.

Reardon, Sean F., and Glenn Firebaugh, "Measures of Multigroup Segregation," *Sociological Methodology*, 32 (2002), 33–67.

Reardon, Sean F., and David O'Sullivan, "Measures of Spatial Segregation," *Sociological Methodology*, 34 (2004), 121–162.

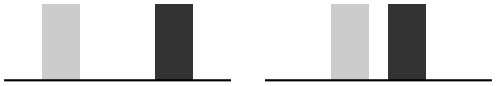
U.S. Geological Survey, *Global Land Cover Characteristics Data Base Version 2.0*, Reston,

VA (2000).

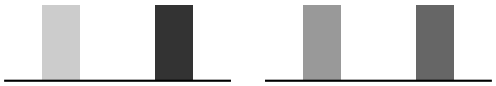
Weidmann, Nils B., "Geography as Motivation and Opportunity: Group Concentration and Ethnic Conflict, *Journal of Conflict Resolution*, 53 (2009), 526–543.

White, Michael J., "The Measurement of Spatial Segregation," *American Journal of Sociology*, 88 (1983), 1008–1018.

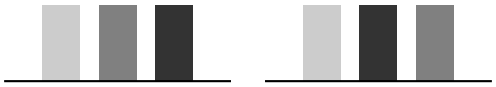
Figures and Tables



(a) Importance of spatial distances



(b) Importance of ethnic distances



(c) Importance of alignment

Figure 1: Illustration of our segregation measure

Notes: The two diagrams of each sub-figure depict two distributions of ethnic groups in space. Each tone of gray indicates a different ethnic group, and ethnic distances between groups are given by differences in tones of gray. Spatial locations are on the horizontal axis, which also measures spatial distances, while the vertical axis measures the population mass at each location.

(a) Distributions of Axiom 1.

(b) Distributions of Axiom 2.

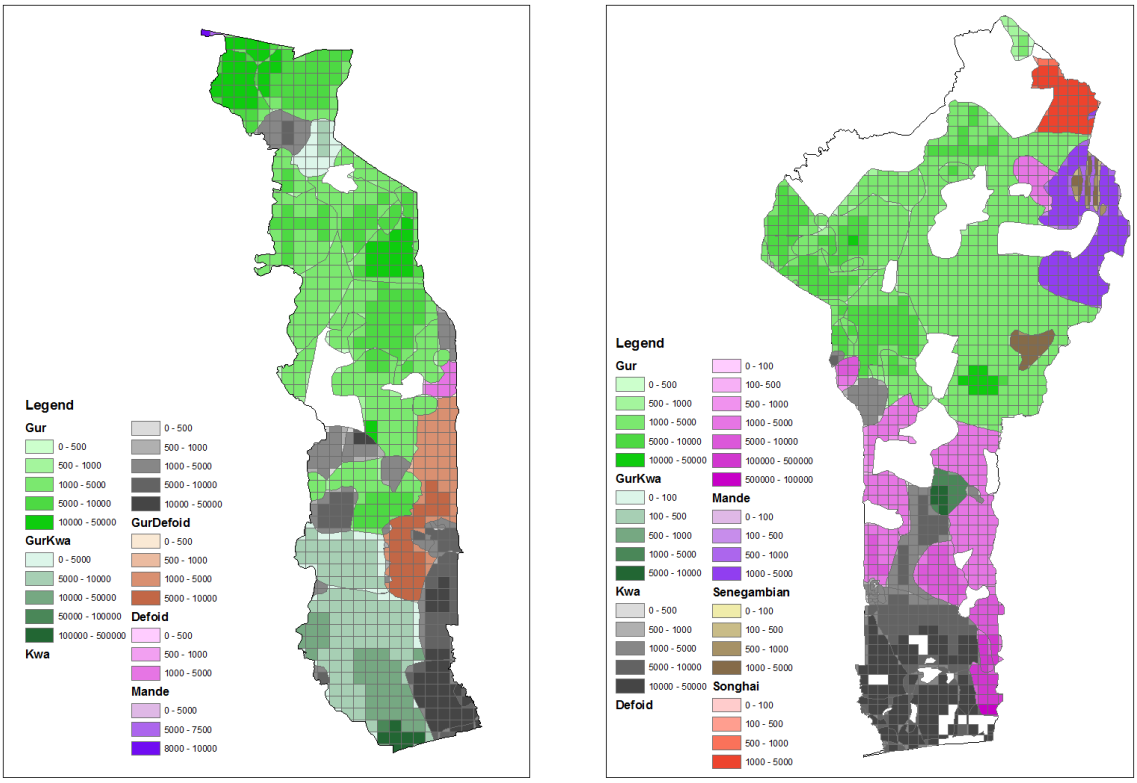
(c) Distributions of Axiom 3.

(d) Distributions of Axiom 4.

Figure 2: Illustration of the distributions of the axiomatization

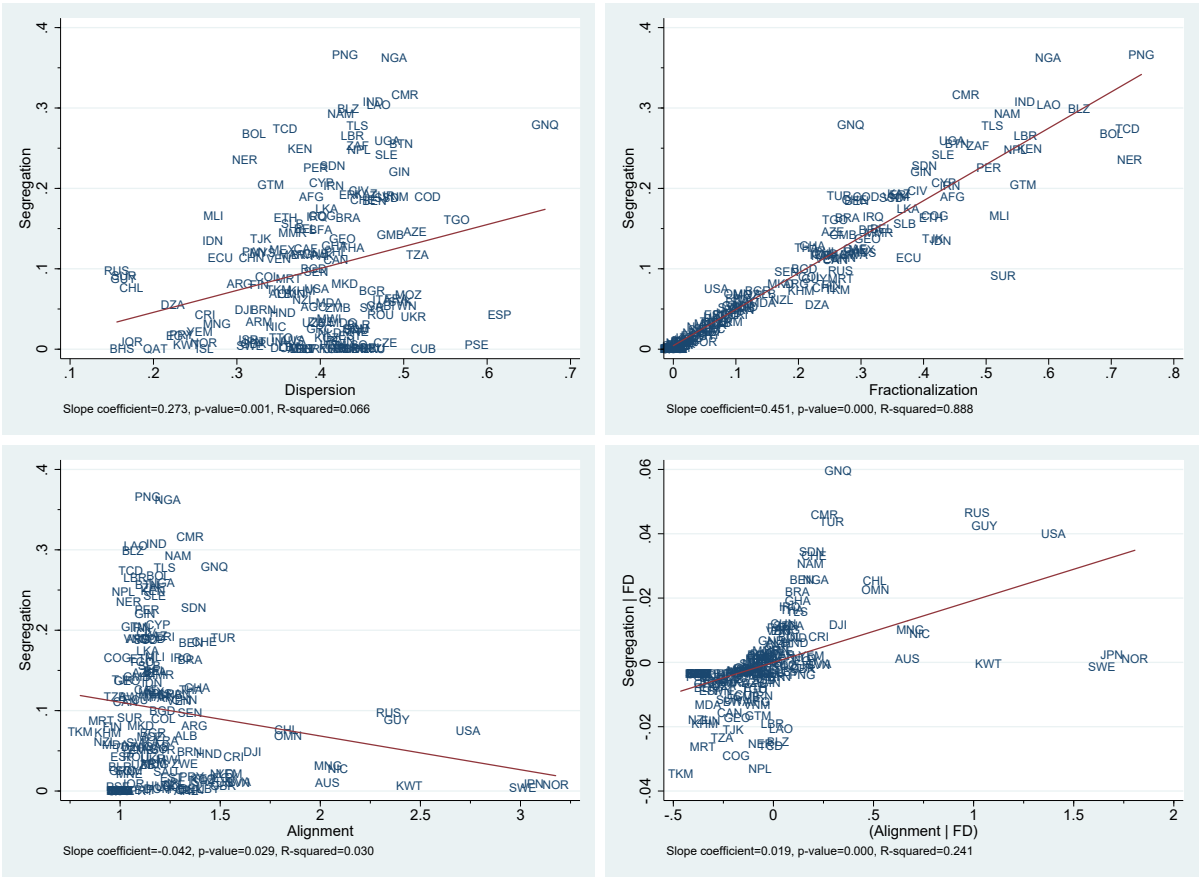
Notes: The two diagrams of each sub-figure depict two distributions of ethnic groups in space. Each tone of gray indicates a different ethnic group, and ethnic distances between groups are given by differences in tones of gray. Spatial locations are on the horizontal axis, which also measures spatial distances, while the vertical axis measures the population mass at each location.

Figure 3: Ethnographic maps and population data for Togo and Benin



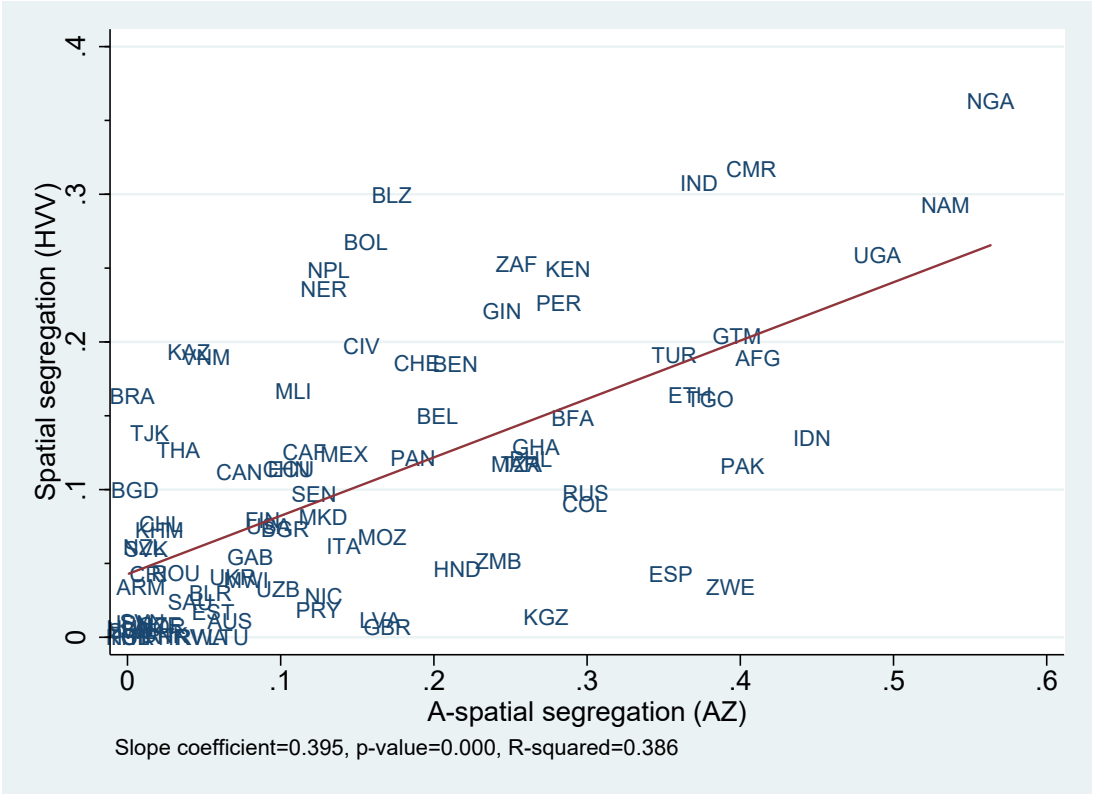
Notes: Maps of Togo (left) and Benin (right) showing the traditional homelands of language groups according to WLMS and our grid cells. Each grid cell constitutes a different location in the computation of our indices, each color indicates that the corresponding grid cell belongs to the traditional homeland of a certain language group (with the relevant language groups given in the legend), and the brightness of this color indicates the current population (also given in the legend). The legend entries Gur/Kwa and Gur/Defoid indicate the traditional homelands of multiple language groups, some speaking a Gur language and some a Kwa or Defoid language. WLMS indicates no traditional homelands in the white areas.

Figure 4: Scatter plots illustrating the index of ethnic segregation and its components



Notes: Scatter plots showing the associations between the index of ethnic segregation S and its three components: spatial dispersion (D, top left), generalized ethnic fractionalization (F, top right) and alignment (A, bottom left). Additional scatter plot showing the association between S and A after partialling out benchmark segregation (i.e., F×D) from both S and A (bottom right). The (red) lines indicate the best linear fit.

Figure 5: Scatter plot comparing indices of ethnic segregation



Notes: The scatter plot shows the association between our spatial index of ethnic segregation (on the vertical axis) and the a-spatial segregation index by Alesina and Zhuravskaya (2011, on the horizontal axis), which is based on the population shares of different ethnolinguistic groups in different subnational units rather than ethnic and spatial distances.

Table 1: Summary statistics for our indices of ethnic geography

	Obs.	Mean	Std. Dev.	Min.	Max.
Segregation	161	0.099	0.093	0	0.366
Alignment	161	1.274	0.383	0.801	3.176
Benchmark Seg.	161	0.084	0.081	0	0.322
Fractionalization	161	0.210	0.195	0	0.748
Dispersion	161	0.395	0.088	0.156	0.669

Table 2: Ethnic geography and the rule of law

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Segregation	-3.93***	-2.08***	-1.39*	-1.52*	-1.93***	-2.14***	-1.24**	-2.13***
	(0.69)	(0.66)	(0.71)	(0.85)	(0.70)	(0.61)	(0.56)	(0.70)
R^2	0.14	0.39	0.44	0.43	0.40	0.44	0.39	0.46
Alignment	0.62***	0.57***	0.52***	0.52**	0.59***	0.50***	0.65*	0.50***
	(0.18)	(0.18)	(0.17)	(0.21)	(0.18)	(0.18)	(0.37)	(0.18)
Benchmark Seg.	-3.90***	-1.95**	-1.26	-1.29	-1.69**	-2.07***	-1.30	-2.08**
	(0.78)	(0.77)	(0.81)	(0.96)	(0.79)	(0.70)	(0.80)	(0.81)
R^2	0.20	0.44	0.48	0.47	0.46	0.48	0.42	0.50
Alignment	0.62***	0.51***	0.47***	0.52**	0.56***	0.46***	0.64*	0.41**
	(0.19)	(0.18)	(0.17)	(0.20)	(0.18)	(0.17)	(0.37)	(0.18)
Fractionalization	-1.63***	-0.81**	-0.53	-0.65	-0.74**	-0.83**	-0.65*	-0.83**
	(0.32)	(0.33)	(0.34)	(0.42)	(0.35)	(0.32)	(0.39)	(0.37)
Dispersion	-0.22	-0.90	-0.76	0.47	-0.42	-0.82	-0.36	-1.33
	(0.85)	(0.85)	(0.89)	(1.04)	(1.03)	(0.83)	(0.80)	(0.87)
R^2	0.20	0.44	0.48	0.47	0.46	0.49	0.42	0.51
Continent FE	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Further controls	No	No	Climate	Terrain	Deep hist.	Col. hist.	No	No
Population data	Current	Current	Current	Current	Current	Current	Land cov.	Current
Restricted sample	No	No	No	No	No	No	No	Yes
Countries	157	157	156	147	150	157	154	131

Notes: Dependent variable is rule of law in 2010 from the World Bank Governance Indicators. Each column presents three OLS regressions. In the upper panel the main explanatory variable is ethnic segregation, in the middle panel these are ethno-spatial alignment and benchmark segregation, and in the lower panel these are ethno-spatial alignment, generalized ethnic fractionalization and spatial dispersion. These indices are explained in Sections 2 and 3.1. They are computed using current population density data in columns (1)–(6) and (8), and global land cover data on habitable areas in column (7). Columns (2)–(8) include continent fixed effects. Additional controls are temperature, precipitation and absolute latitude in column (3); terrain ruggedness, its interaction with a dummy variable for Africa, and averages and standard deviations of elevation and land suitability for agriculture in column (4); migratory distance from Addis Ababa, its square term, and the time elapsed since the agricultural transition in column (5); and dummy variables for former British/French/Spanish/other colonies in column (6). Online Appendix E contains more information on dependent and control variables. Settler colonies, defined as former colonies where more than 10 percent of the current population has ancestors from former European colonial powers, according to Putterman and Weil's (2010) world migration matrix, are excluded in column (8). Robust standard errors. ***, **, * indicate p-values below 0.01, 0.05 and 0.1, respectively.

Table 3: Ethnic geography and income

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Segregation	-5.19***	-1.63*	-0.86	-0.72	-2.06**	-1.85**	-1.10*	-2.42**
	(1.00)	(0.98)	(0.99)	(1.13)	(0.96)	(0.93)	(0.65)	(1.16)
R^2	0.14	0.52	0.55	0.61	0.54	0.55	0.54	0.55
Alignment	0.69***	0.55***	0.56***	0.37**	0.55***	0.50***	0.85***	0.57***
	(0.17)	(0.16)	(0.16)	(0.19)	(0.16)	(0.15)	(0.32)	(0.17)
Benchmark Seg.	-5.56***	-1.84*	-1.07	-0.99	-2.09*	-2.08**	-1.25	-2.78**
	(1.15)	(1.09)	(1.09)	(1.24)	(1.09)	(1.04)	(0.80)	(1.27)
R^2	0.21	0.55	0.58	0.62	0.57	0.58	0.59	0.58
Alignment	0.61***	0.49**	0.52***	0.38*	0.44**	0.44**	0.93***	0.47**
	(0.18)	(0.19)	(0.18)	(0.19)	(0.19)	(0.18)	(0.31)	(0.22)
Fractionalization	-2.19***	-0.78*	-0.53	-0.72	-0.80*	-0.80*	-0.41	-1.17**
	(0.46)	(0.41)	(0.44)	(0.49)	(0.43)	(0.41)	(0.37)	(0.48)
Dispersion	-1.61	-1.04	-0.43	1.68	-1.69	-1.06	-2.42***	-1.39
	(1.13)	(1.22)	(1.20)	(1.25)	(1.08)	(1.18)	(0.91)	(1.40)
R^2	0.21	0.56	0.58	0.64	0.58	0.58	0.61	0.59
Continent FE	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Further controls	No	No	Climate	Terrain	Deep hist.	Col. hist.	No	No
Population data	Current	Current	Current	Current	Current	Current	Land cov.	Current
Restricted sample	No	No	No	No	No	No	No	Yes
Countries	148	148	148	139	144	148	145	123

Notes: Dependent variable is log of expenditure-side real GDP per capita in 2010 from the Penn World Tables 9.0. Each column presents three OLS regressions. In the upper panel the main explanatory variable is ethnic segregation, in the middle panel these are ethno-spatial alignment and benchmark segregation, and in the lower panel these are ethno-spatial alignment, generalized ethnic fractionalization and spatial dispersion. These indices are explained in Sections 2 and 3.1. They are computed using current population density data in columns (1)–(6) and (8), and global land cover data on habitable areas in column (7). Columns (2)–(8) include continent fixed effects. Additional controls are temperature, precipitation and absolute latitude in column (3); terrain ruggedness, its interaction with a dummy variable for Africa, and averages and standard deviations of elevation and land suitability for agriculture in column (4); migratory distance from Addis Ababa, its square term, and the time elapsed since the agricultural transition in column (5); and dummy variables for former British/French/Spanish/other colonies in column (6). Online Appendix E contains more information on dependent and control variables. Settler colonies, defined as former colonies where more than 10 percent of the current population has ancestors from former European colonial powers, according to Putterman and Weil's (2010) world migration matrix, are excluded in column (8). Robust standard errors. ***, **, * indicate p-values below 0.01, 0.05 and 0.1, respectively.

Table 4: Ethnic geography and trust

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Segregation	-0.21	-0.00	0.20	-0.02	0.03	0.04	0.11	-0.04
	(0.21)	(0.19)	(0.16)	(0.18)	(0.18)	(0.19)	(0.13)	(0.22)
R^2	0.01	0.25	0.40	0.48	0.35	0.28	0.26	0.23
Alignment	0.12***	0.12***	0.10***	0.07**	0.10***	0.11***	0.14***	0.13***
	(0.03)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.05)	(0.04)
Benchmark Seg.	-0.02	0.22	0.36*	0.21	0.22	0.24	0.17	0.19
	(0.23)	(0.21)	(0.19)	(0.21)	(0.21)	(0.21)	(0.19)	(0.25)
R^2	0.15	0.38	0.48	0.53	0.45	0.40	0.28	0.39
Alignment	0.12***	0.10***	0.09***	0.08**	0.09**	0.10**	0.15***	0.11***
	(0.04)	(0.03)	(0.03)	(0.03)	(0.04)	(0.04)	(0.06)	(0.04)
Fractionalization	-0.02	0.10	0.16*	0.07	0.09	0.11	0.11	0.11
	(0.09)	(0.09)	(0.08)	(0.09)	(0.09)	(0.10)	(0.10)	(0.11)
Dispersion	0.03	-0.16	-0.07	0.12	-0.14	-0.19	-0.12	-0.24
	(0.19)	(0.19)	(0.20)	(0.21)	(0.18)	(0.22)	(0.15)	(0.21)
R^2	0.15	0.39	0.49	0.53	0.46	0.41	0.29	0.41
Continent FE	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Further controls	No	No	Climate	Terrain	Deep hist.	Col. hist.	No	No
Population data	Current	Current	Current	Current	Current	Current	Land cov.	Current
Restricted sample	No	No	No	No	No	No	No	Yes
Countries	77	77	77	74	76	77	77	63

Notes: Dependent variable is generalized trust from the World Value Survey in the 1981-2008 time period (taken from Ashraf and Galor 2013). It is the fraction of people answering "most people can be trusted" (as opposed to "can't be too careful") when asked the standard trust question. Each column presents three OLS regressions. In the upper panel the main explanatory variable is ethnic segregation, in the middle panel these are ethno-spatial alignment and benchmark segregation, and in the lower panel these are ethno-spatial alignment, generalized ethnic fractionalization and spatial dispersion. These indices are explained in Sections 2 and 3.1. They are computed using current population density data in columns (1)–(6) and (8), and global land cover data on habitable areas in column (7). Columns (2)–(8) include continent fixed effects. Additional controls are temperature, precipitation and absolute latitude in column (3); terrain ruggedness, its interaction with a dummy variable for Africa, and averages and standard deviations of elevation and land suitability for agriculture in column (4); migratory distance from Addis Ababa, its square term, and the time elapsed since the agricultural transition in column (5); and dummy variables for former British/French/Spanish/other colonies in column (6). Online Appendix E contains more information on dependent and control variables. Settler colonies, defined as former colonies where more than 10 percent of the current population has ancestors from former European colonial powers, according to Putterman and Weil's (2010) world migration matrix, are excluded in column (8). Robust standard errors. ***, **, * indicate p-values below 0.01, 0.05 and 0.1, respectively.

Online Appendix to "Ethnic Geography: Measurement and Evidence"

Roland Hodler, Michele Valsecchi and Alberto Vesperoni[1]

Sections:

[1]Hodler: Department of Economics, University of St.Gallen; CEPR, London; CESifo, Munich; email: roland.hodler@unisg.ch.
Valsecchi: New Economic School, Moscow; email: mvalsecchi@nes.ru.
Vesperoni: Department of Economics, University of Klagenfurt; email: alberto.vesperoni@gmail.com.

A. Shortcomings of a-spatial segregation indices

Border dependence: Border dependence occurs due to the (implicit) assumption of a-spatial segregation measures that the distance between two individuals is zero when they are located in the same subnational unit, and one when located in different subnational units. As a result, the index value of a-spatial segregation measures heavily depends on the type of subnational units used when computing the index values. For example, it may depend on whether provinces or districts are used when relying on administrative units, or on the size of cells or circles when researchers construct "geometric" subnational units.

Figure A.1 illustrates the problem of border dependence: The spatial distribution of individuals from different ethnic groups is identical in the left and the right diagram, however there are four administrative units in the left diagram, but only two in the right diagram. Any a-spatial segregation measure would classify the society in the left diagram as highly segregated, because the population is ethnically homogenous in each administrative unit, but as non-segregated in the right diagram, where the two groups' population shares are the same in each administrative unit.

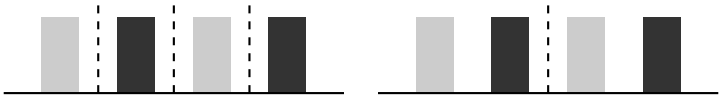


Figure A.1: Illustration of border dependence
Notes: The two diagrams depict two distributions of ethnic groups in space. Each tone of gray indicates a different ethnic group, and ethnic distances between groups are given by differences in tones of gray. Spatial locations are on the horizontal axis, which also measures spatial distances, while the vertical axis measures the population mass at each location. The dotted vertical lines indicate administrative boundaries.

To illustrate that border dependence is a real concern, we use data from the Nigeria Development and Health Survey (DHS) 2013. This survey of more than 38,000 mothers of childbearing age provides information on, among other things, these mothers' self-reported ethnicity and the geo-coordinates of cluster locations. We use these geo-coordinates to assign each cluster (and thereby each mother) to a state and a local government area (LGA). The DHS further groups Nigeria into 6 regions that play no administrative or political role. Table A.1, column (1) shows that, according to the Nigeria DHS 2013, there are 307 different ethnic groups and the population share of the largest group (Hausa) is 24 percent. We then collapse the data at the level of DHS regions, states and LGAs. For each of these levels, we report in columns (2)–(4) the average number of groups, the average population share of the largest group, and the number of subnational units on which these two summary statistics are based. We see an inverse relation between the level of spatial disaggregation and the average ethnic heterogeneity within subnational units. As a result, any a-spatial segregation index would provide markedly different index values for Nigeria in 2013, depending on whether DHS regions, states or LGAs were used

as the relevant subnational units. The index value would be highest for LGAs and lowest for DHS regions.[2]

Table A.1: Ethnic heterogeneity in subnational units in Nigeria

	(1)	(2)	(3)	(4)
	Country	DHS regions	States	LGAs
Number of units	1	6	38	501
Average number of groups	307	98.17	28.29	5.08
Average share of largest group	0.24	0.53	0.59	0.80

Checkerboard problem: The checkerboard problem refers to the impossibility of a-spatial segregation measures to account for the arrangements or relative positions of subnational units in space. It occurs due to the (implicit) assumption of a-spatial segregation measures that the distance between two individuals is one when they are located in different subnational units, no matter how far apart these units are.

Figure A.2 illustrates the problem: A-spatial segregation measures classify the societies in the left and the right diagram as equally segregated, even though the society represented in the left diagram appears more segregated than the one in the right diagram.

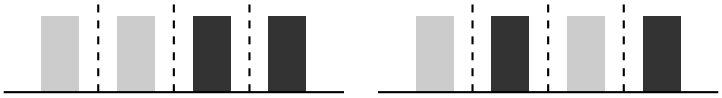


Figure A.2: Illustration of the checkerboard problem
Notes: The two diagrams of each sub-figure depict two distributions of ethnic groups in space. Each tone of gray indicates a different ethnic group, and ethnic distances between groups are given by differences in tones of gray. Spatial locations are on the horizontal axis, which also measures spatial distances, while the vertical axis measures the population mass at each location. The dotted vertical lines indicate administrative boundaries.

[2]Alesina and Zhursavskaysa (2011) use DHS to compute ethnic segregation in various countries, including Nigeria, where they take DHS regions as the relevant subnational units.

B. Geometric interpretation of our segregation index

To illustrate the general properties of our segregation index and its various components, we now provide a geometric interpretation. Suppose the population is finite, where $P := \{1, \ldots, m\}$ is the set of individuals and $m \geq 3$. For each pair of individuals $i, j \in P$, denote by $\lambda_{i,j}$ and $\gamma^{i,j}$ the spatial and ethnic distance between them. Let

$$\Lambda := (\lambda_{1,1}, \ldots, \lambda_{m,m}) \text{ and } \Gamma := (\gamma^{1,1}, \ldots, \gamma^{m,m})$$

be the vectors of spatial and ethnic distances between all unordered pairs of individuals. Then, equation (2) can be written as $S(\mu, \lambda, \gamma) = \frac{4}{m^2}\Lambda \cdot \Gamma$, and by definition of inner product our segregation index can be decomposed into

$$S(\mu, \lambda, \gamma) = \frac{4}{m^2}||\Lambda||_2||\Gamma||_2 \cos[\theta_{\Lambda,\Gamma}], \tag{B.1}$$

where

$$||\Lambda||_2 := \left(\frac{1}{2}\sum_{(i,j)\in P^2}(\lambda_{i,j})^2\right)^{1/2} \text{ and } ||\Gamma||_2 := \left(\frac{1}{2}\sum_{(i,j)\in P^2}(\gamma^{i,j})^2\right)^{1/2}$$

are the Euclidean norms of the two vectors Λ and Γ, and $\theta_{\Lambda,\Gamma}$ is the angle between them.

Since $\cos[0] = 1$, our segregation index is maximized when the two vectors point in the same direction ($\theta_{\Lambda,\Gamma} = 0$), which means that Λ and Γ are linearly dependent, i.e., there is some $k > 0$ such that $\lambda_{i,j} = k\gamma^{i,j}$ for all $i, j \in P$. In this sense, S can be interpreted as a geometric projection. To see an example, consider the two joint distributions in Figure 1(c). Clearly, by S the left distribution is more segregated than the right, as Λ and Γ are co-directional in the left but not in the right distribution, everything else equal. This is in line with our intuition in the Introduction. Another relevant feature of our index is that any increase in the mean of the two vectors, or in their Euclidean norms, also leads to higher segregation. For example, in Figure 1(b) the distribution on the left is more segregated than that on the right as the mean ethnic distance (and the Euclidean norm $||\Gamma||_2$) is higher, everything else being equal. Moreover, any mean-preserving spread of the elements of each of the two vectors Λ and Γ that keeps their alignment constant leads to higher segregation. This can be easily shown by the convexity of the (square of the) Euclidean norms $||\Lambda||_2$ and $||\Gamma||_2$ in the spatial distance and in the ethnic distance between each pair of individuals, respectively.

This geometric interpretation of our segregation index resembles the decomposition in Proposition 1: The generalized social fractionalization index F and the spatial dispersion index D are related to the Euclidean norms of the two respective vectors, and the alignment index A is therefore related to the cosign of the angle between the vectors of ethnic and spatial distances. In particular, it follows from Proposition 1 and Equation

(B.1) that $A(\mu, \lambda, \gamma) \approx \cos[\theta_{\Lambda,\Gamma}]$ and $F(\mu, \gamma)D(\mu, \lambda) \approx 4||\Lambda||_2||\Gamma||_2/m^2$. To see this, it is useful to write

$$F(\mu, \gamma)D(\mu, \lambda) = \left(\frac{2}{m^2}\right)^2 \left(\sum_{(i,j)\in P^2} \gamma^{i,j}\right) \left(\sum_{(i,j)\in P^2} \lambda_{i,j}\right),$$

$$4||\Lambda||_2||\Gamma||_2/m^2 = \left(\frac{2}{m^2}\right) \left(\sum_{(i,j)\in P^2} (\gamma^{i,j})^2\right)^{1/2} \left(\sum_{(i,j)\in P^2} (\lambda_{i,j})^2\right)^{1/2}.$$

Note the proportionality across the two equations for each of the three elements that respectively correspond to population size (m), social distances ($\gamma^{i,j}$) and spatial distances ($\lambda_{i,j}$). Although different, $F(\mu, \gamma)D(\mu, \lambda)$ and $4||\Lambda||_2||\Gamma||_2/m^2$ are closely related, which means that $A(\mu, \lambda, \gamma)$ and the cosign of $\theta_{\Lambda,\Gamma}$ are closely related as well.[3] This relation further justifies our interpretation of A as alignment or co-directionality of spatial and ethnic distances. For the purpose of empirical applications, A has the advantage – compared to the consign of $\theta_{\Lambda,\Gamma}$ – that its computation does not require data at the individual level. Similarly, F and D are related to the Euclidean norms $||\Gamma||_2$ and $||\Lambda||_2$ and have the same empirical advantage compared to them.

[3]One can show that $A(\mu, \lambda, \gamma)$ is a positively-biased proxy of $\cos[\theta_{\Lambda,\Gamma}]$. This follows from $4||\Lambda||_2||\Gamma||_2/m^2 \geq S(\mu, \lambda, \gamma)$ for all $\mu \in \mathcal{M}$ (as $\cos[\theta_{\Lambda,\Gamma}] \in [0, 1]$) and $F(\mu, \gamma)D(\mu, \lambda) = S(\overline{\mu}, \lambda, \gamma)$, which jointly imply $4||\Lambda||_2||\Gamma||_2/m^2 \geq F(\mu, \gamma)D(\mu, \lambda)$. Hence, $A(\mu, \lambda, \gamma) \geq \cos[\theta_{\Lambda,\Gamma}]$.

C. List of countries

We provide our four indices of ethnic geography (i.e., ethnic segregation, generalized ethnic fractionalization, spatial dispersion, and ethno-spatial alignment) for the following 161 countries with a current population of more than 250,000 and a land surface area of more than 5,000 km^2: Afghanistan, Albania, Algeria, Angola, Argentina, Armenia, Australia, Azerbaijan, Bahamas, Bangladesh, Belarus, Belgium, Belize, Benin, Bhutan, Bolivia, Bosnia and Herzegovina, Botswana, Brazil, Brunei, Bulgaria, Burkina Faso, Burundi, Cambodia, Cameroon, Canada, Central African Republic, Chad, Chile, China, Colombia, Congo, Costa Rica, Cote d'Ivoire, Croatia, Cuba, Cyprus, Czech Republic, Democratic Republic of the Congo, Denmark, Djibouti, Dominican Republic, East Timor, Ecuador, Egypt, El Salvador, Equatorial Guinea, Eritrea, Estonia, Ethiopia, Finland, France, Gabon, Gambia, Georgia, Germany, Ghana, Greece, Guatemala, Guinea, Guinea-Bissau, Guyana, Haiti, Honduras, Hungary, Iceland, India, Indonesia, Iran, Iraq, Ireland, Israel, Italy, Jamaica, Japan, Jordan, Kazakhstan, Kenya, Kuwait, Kyrgyzstan, Laos, Latvia, Lebanon, Lesotho, Liberia, Libya, Lithuania, Macedonia, Madagascar, Malawi, Malaysia, Mali, Mauritania, Mexico, Moldova, Mongolia, Montenegro, Morocco, Mozambique, Myanmar, Namibia, Nepal, Netherlands, New Zealand, Nicaragua, Niger, Nigeria, Norway, Oman, Pakistan, Palestine, Panama, Papua New Guinea, Paraguay, Peru, Philippines, Poland, Portugal, Qatar, Romania, Russia, Rwanda, Saudi Arabia, Senegal, Sierra Leone, Slovakia, Slovenia, Solomon Islands, Somalia, South Africa, South Korea, South Sudan, Spain, Sri Lanka, Sudan, Suriname, Swaziland, Sweden, Switzerland, Syria, Taiwan, Tajikistan, Tanzania, Thailand, Togo, Trinidad and Tobago, Tunisia, Turkey, Turkmenistan, Uganda, Ukraine, United Arab Emirates, United Kingdom, United States, Uruguay, Uzbekistan, Venezuela, Vietnam, Yemen, Zambia, Zimbabwe.

D. Correlations between our indices and alternative indices

Table D.1: Correlations between our indices and alternative indices

Index (i)	Source	Cor(S,i)	Cor(A,i)	Cor(F,i)	Cor(D,i)	Obs.
Standard fractionalization	ADEKW	0.578	-0.235	0.541	0.304	154
Standard fractionalization	AZ	0.581	-0.223	0.554	0.246	91
A-spatial segregation	AZ	0.621	-0.129	0.541	0.164	90
Standard fractionalization	EMR	0.585	-0.216	0.569	0.115	133
Generalized fractionalization	EMR	0.597	-0.081	0.631	-0.033	133
Polarization	EMR	0.384	-0.049	0.441	-0.080	133

Notes: Standard fractionalization refers to the index of ethnic fractionalization based on categorical data, whereas generalized fractionalization is based on (non-binary) ethnic distances and sometimes called the Greenberg-Gini index. A-spatial segregation refers to the segregation index used by Alesina and Zhuravskaya (2011), which is based on the population shares of different ethnic groups in different subnational units rather than ethnic and spatial distances. Polarization refers to the polarization index by Duclos et al. (2004). ADEKW stands for Alesina et al. (2003), AZ for Alesina and Zhuravskaya (2011), and EMR for Esteban et al. (2012). Cor(X,i) refers to the correlation between our index X and the index i given in the first column.

E. Definitions of dependent and control variables

E.1. Dependent variables

E.1.1. Main dependent variables

Rule of law: This is one of six World Bank Governance Indicators (also called Worldwide Governance Indicators) for 2010. These indicators are based on several hundred individual variables from many different organizations measuring perceptions of governance. These individual measures of governance are assigned to categories capturing key dimensions of governance. An unobserved component model is used to construct the six aggregate governance indicators. They are normally distributed with a mean of zero and a standard deviation of one each year of measurement. The rule of law indicator includes several indicators that measure the extent to which agents have confidence in and abide by the rules of society. These include perceptions of the incidence of crime, the effectiveness and predictability of the judiciary, and the enforceability of contracts. This indicator thus measures the success of a society in developing an environment in which fair and predictable rules form the basis for economic and social interactions and the extent to which property rights are protected.

Income (PWT): Logarithm of expenditure-side real GDP per capita in 2010 at chained purchasing power parities (in 2011 US dollars) by Penn World Table, version 9.

Trust: Measure of generalized trust based on World Values Surveys conducted from 1981-2008. It is calculated as the fraction of total respondents who responded with "most people can be trusted" (as opposed to "can't be too careful") when asked: "Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?" Variable taken from Ashraf and Galor (2013).

E.1.2. Additional dependent variables used in Online Appendix E

Control of corruption: This is one of six World Bank Governance Indicators for 2010. It measures perceptions of corruption, including the frequency of bribe payments in the business environment and the extent of political corruption.

Government effectiveness: This is one of six World Bank Governance Indicators for 2010. It measures public service provision, the quality of the bureaucracy, the competence of civil servants, and the independence of the civil service from political pressures.

Political stability: This is one of six World Bank Governance Indicators for 2010. It measures perceptions of the likelihood that the government in power will be destabilized

or overthrown by possibly unconstitutional and/or violent means.

Regulatory quality: This is one of six World Bank Governance Indicators for 2010. It measures the incidence of market-unfriendly policies and perceptions of the burdens imposed by excessive regulation in areas such as foreign trade and business development.

Voice and accountability: This is one of six World Bank Governance Indicators for 2010. It measures various aspects of the political process, civil liberties and political rights to indicate the extent to which citizens of a country are able to participate in the selection of governments.

Quality of government: This indicator from the International Country Risk Guide (ICRG) corresponds to the mean of three ICRG variables in 2010: Corruption, law and order, and bureaucratic quality.

Corruption perception index: This index from Transparency International focuses on perceptions of corruption in the public sector in 2010 and includes both administrative and political corruption. We have rescaled it so that it ranges between zero and one, with higher values implying less corruption.

Income (WDI): Logarithm of GDP per capita in 2010 based on purchasing power parity (in constant 2011 international dollars) from the World Development Indicators.

E.1.3. Summary statistics

Table E.1: Summary statistics for our dependent variables

	Observations	Mean	Std. Dev.	Min.	Max.
Rule of law	157	-0.203	0.988	-2.448	1.977
Income (PWT, in logs)	148	9.047	1.241	6.341	11.708
Trust	77	0.277	0.141	0.038	0.664
Control of corruption	157	-0.171	0.988	-1.739	2.414
Government effectiveness	157	-0.119	0.979	-2.239	2.245
Political stability	157	0.255	0.384	0.000	1.393
Regulatory quality	157	-0.096	0.976	-2.381	1.888
Voice and accountability	157	-0.213	0.995	-2.155	1.637
Quality of government	131	0.526	0.198	0.083	1.000
Corruption perception index	154	0.385	0.205	0.110	0.930
Income (WDI, in logs)	152	9.040	1.256	6.391	11.757

E.2. Control variables

Absolute latitude: The absolute value of the latitude of a country's approximate centroid, as reported by the CIA's World Factbook, taken from Ashraf and Galor (2013).

Temperature: The intertemporal average monthly temperature of a country in degrees Celsius per month over the 1961–1990 time period, calculated using geospatial average monthly temperature data, taken from Ashraf and Galor (2013).

Precipitation: The intertemporal average monthly precipitation of a country in mm per month over the 1961–1990 time, calculated using geospatial average monthly precipitation data, taken from Ashraf and Galor (2013).

Terrain roughness: Terrain Ruggedness Index by Nunn and Puga (2012), which quantifies average local topographic heterogeneity by measuring elevation differences for grid points within 30 arc-seconds.

Average and standard deviation of elevation: Variables based on geospatial elevation data, taken from Michalopoulos (2012).

Average and standard deviation of land suitability: Variables based on a geospatial index of the suitability of land for agriculture based on ecological indicators of climate and soil suitability for cultivation, taken from Michalopoulos (2012).

Migratory distance from Addis Ababa: The great circle distance from Addis Ababa (Ethiopia) to the country's modern capital city along a land-restricted path forced through one or more of five intercontinental waypoints (Cairo, Istanbul, Phnom Penh, Anadyr, and Prince Rupert), taken from Ashraf and Galor (2013).

Time elapsed since the agricultural transition: The number of years elapsed up to the year 2000 CE since the majority of the population residing within a country's modern national borders began practicing sedentary agriculture as the primary mode of subsistence, taken from Ashraf and Galor (2013).

Former colonizer: A variable indicating whether a country is a former British colony, a former French colony, a former Spanish colony, the former colony of another Western colonizer, or not a former Western colony. It is based on the classification of Western overseas colonies in the Authoritarian Regime Dataset.

F. Trust as a possible mechanism

Table F.1 Ethnic geography, trust, rule of law, and income

	(1)	(2)	(3)	(4)
Dependent var.	Rule of law	Rule of law	Income	Income
Alignment	0.60***	0.28	0.48***	0.29*
	(0.19)	(0.21)	(0.14)	(0.15)
Fractionalization	-0.66	-0.97	-0.50	-0.69
	(0.72)	(0.70)	(0.65)	(0.64)
Dispersion	0.10	0.60	-0.34	-0.04
	(1.46)	(1.27)	(1.01)	(0.98)
Trust		3.02***		1.81***
		(0.85)		(0.62)
Continent FE	Yes	Yes	Yes	Yes
Countries	77	77	77	77
R^2	0.42	0.52	0.59	0.63

Notes: OLS regressions with continent fixed effects. The dependent variable is the rule of law in 2010 from the World Bank Governance Indicators in columns (1) and (2), and expenditure-side real GDP per capita in 2010 from the Penn World Tables 9.0 in columns (3) and (4). The sample is restricted to countries for which generalized trust from the World Value Survey in the 1981-2008 time period is available. Online Appendix E contains more information on the dependent variables and generalized trust. Ethno-spatial alignment, generalized ethnic fractionalization and spatial dispersion are explained in Sections 2 and 3.1. Robust standard errors. ***, **, * indicate p-values below 0.01, 0.05 and 0.1, respectively.

G. Robustness of cross-country regressions

Table G.1: Ethnic geography and the rule of law in restricted samples

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent var.	Rule of law (WBGI)							
Segregation	-2.94***	-2.31***	-1.54**	-2.17***	-1.66**	-1.68***	-2.09***	-1.90***
	(0.93)	(0.75)	(0.76)	(0.67)	(0.65)	(0.63)	(0.73)	(0.62)
R^2	0.33	0.45	0.44	0.18	0.38	0.42	0.39	0.46
Alignment	0.55***	0.56***	0.51**	0.63***	0.58***	0.53***	0.65***	0.62***
	(0.19)	(0.18)	(0.23)	(0.22)	(0.18)	(0.18)	(0.19)	(0.13)
Benchmark Seg.	-2.63**	-2.22**	-1.47*	-2.06**	-1.50**	-1.60**	-1.45*	-1.69**
	(1.07)	(0.88)	(0.88)	(0.79)	(0.75)	(0.73)	(0.87)	(0.71)
R^2	0.39	0.49	0.47	0.25	0.44	0.46	0.46	0.50
Alignment	0.52***	0.50***	0.45*	0.57**	0.52***	0.43**	0.64***	0.58***
	(0.19)	(0.18)	(0.24)	(0.23)	(0.18)	(0.18)	(0.19)	(0.14)
Fractionalization	-1.19**	-0.94**	-0.60*	-0.77**	-0.63*	-0.62*	-0.63	-0.48
	(0.45)	(0.41)	(0.36)	(0.32)	(0.33)	(0.32)	(0.40)	(0.30)
Dispersion	-0.48	-0.91	-0.91	-1.19	-0.88	-1.31	-0.10	-1.17*
	(1.21)	(0.96)	(1.02)	(0.74)	(0.86)	(0.81)	(0.95)	(0.67)
R^2	0.40	0.50	0.48	0.26	0.44	0.47	0.46	0.51
Continent FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Omitted observations	Africa	Americas	Asia	Europe	Oceania	Neo-Europe	$F = 0$	Outliers
Countries	109	129	115	122	153	153	142	148/149/147

Notes: Dependent variable is the rule of law in 2010 by the World Bank Governance Indicators. Each column presents three OLS regressions with continent fixed effects. We omit countries from one continent in each of the columns (1)–(5), Australia, Canada, New Zealand and United States in column (6), the ethnically homogeneous countries in column (7), and outliers as identified by Cook's distance (with a threshold of 4/157) in column (8). Segregation, alignment, benchmark segregation, fractionalization and dispersion are explained in Sections 2 and 3.1. Robust standard errors. ***, **, * indicate p-values below 0.01, 0.05 and 0.1, respectively.

Table G.2: Ethnic geography and income in restricted samples

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent var.	Income (PWT)							
Segregation	-2.88***	-1.95*	0.04	-1.86*	-1.62*	-1.63*	-1.77*	-1.97**
	(1.09)	(1.16)	(1.10)	(1.06)	(0.98)	(0.97)	(1.03)	(0.90)
R^2	0.26	0.55	0.62	0.40	0.51	0.51	0.53	0.57
Alignment	0.45***	0.56***	0.53***	0.68***	0.56***	0.53***	0.60***	0.58***
	(0.16)	(0.17)	(0.18)	(0.25)	(0.17)	(0.17)	(0.16)	(0.14)
Benchmark Seg.	-2.90**	-2.30*	-0.10	-2.01*	-1.84*	-1.91*	-1.60	-2.09**
	(1.27)	(1.30)	(1.22)	(1.18)	(1.08)	(1.08)	(1.13)	(1.00)
R^2	0.32	0.58	0.64	0.45	0.54	0.54	0.58	0.61
Alignment	0.39**	0.49**	0.49**	0.62**	0.50**	0.44**	0.58***	0.56***
	(0.18)	(0.21)	(0.22)	(0.26)	(0.19)	(0.21)	(0.18)	(0.14)
Fractionalization	-1.08**	-1.07**	-0.16	-0.79*	-0.78*	-0.78*	-0.75*	-0.81**
	(0.50)	(0.51)	(0.45)	(0.45)	(0.41)	(0.41)	(0.44)	(0.39)
Dispersion	-1.24	-0.83	-0.49	-1.47	-1.03	-1.32	0.01	-0.61
	(1.14)	(1.44)	(1.45)	(1.46)	(1.21)	(1.27)	(1.19)	(0.77)
R^2	0.32	0.59	0.64	0.46	0.55	0.55	0.58	0.63
Continent FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Omitted observations	Africa	Americas	Asia	Europe	Oceania	Neo-Europe	$F = 0$	Outliers
Countries	103	122	108	113	146	144	134	145/144/143

Notes: Dependent variable is the log of expenditure-side real GDP per capita in 2010 from the Penn World Tables 9.0. Each column presents three OLS regressions with continent fixed effects. We omit countries from one continent in each of the columns (1)–(5), Australia, Canada, New Zealand and United States in column (6), the ethnically homogeneous countries in column (7), and outliers as identified by Cook's distance (with a threshold of 4/148) in column (8). Segregation, alignment, benchmark segregation, fractionalization and dispersion are explained in Sections 2 and 3.1. Robust standard errors. ***, **, * indicate p-values below 0.01, 0.05 and 0.1, respectively.

Table G.3: Ethnic geography and trust in restricted samples

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent var.	Trust (WVS)							
Segregation	-0.03	-0.00	-0.14	0.16	-0.00	-0.02	-0.05	-0.02
	(0.23)	(0.20)	(0.25)	(0.18)	(0.19)	(0.18)	(0.22)	(0.18)
R^2	0.14	0.24	0.28	0.41	0.21	0.27	0.23	0.28
Alignment	0.11***	0.12***	0.13***	0.07*	0.12***	0.12***	0.13***	0.08**
	(0.03)	(0.04)	(0.03)	(0.04)	(0.03)	(0.04)	(0.03)	(0.03)
Benchmark Seg.	0.23	0.22	0.06	0.30	0.20	0.17	0.31	0.25
	(0.26)	(0.23)	(0.28)	(0.20)	(0.21)	(0.21)	(0.24)	(0.18)
R^2	0.30	0.38	0.45	0.45	0.36	0.40	0.37	0.32
Alignment	0.11***	0.10***	0.12***	0.07	0.11***	0.10**	0.12***	0.09**
	(0.03)	(0.04)	(0.04)	(0.04)	(0.03)	(0.04)	(0.04)	(0.04)
Fractionalization	0.10	0.13	0.01	0.18**	0.10	0.09	0.15	0.09
	(0.11)	(0.11)	(0.11)	(0.08)	(0.09)	(0.09)	(0.11)	(0.08)
Dispersion	-0.11	-0.25	-0.08	-0.22	-0.15	-0.23	-0.10	-0.18
	(0.22)	(0.21)	(0.22)	(0.21)	(0.19)	(0.19)	(0.20)	(0.16)
R^2	0.30	0.41	0.45	0.48	0.37	0.42	0.38	0.35
Continent FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Omitted observations	Africa	Americas	Asia	Europe	Oceania	Neo-Europe	$F = 0$	Outliers
Countries	67	65	59	42	75	73	70	76/71/70

Notes: Dependent variable is generalized trust from the World Value Survey in the 1981-2008 time period (Ashraf and Galor 2013). Each column presents three OLS regressions with continent fixed effects. We omit countries from one continent in each of the columns (1)–(5), Australia, Canada, New Zealand and United States in column (6), the ethnically homogeneous countries in column (7), and outliers as identified by Cook's distance (with a threshold of 4/77) in column (8). Segregation, alignment, benchmark segregation, fractionalization and dispersion are explained in Sections 2 and 3.1. Robust standard errors. ***, **, * indicate p-values below 0.01, 0.05 and 0.1, respectively.

Table G.4: Alternative measures of the quality of government and incomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent var.	CC (WBGI)	GE (WBGI)	PS (WBGI)	RQ (WBGI)	V&A (WBGI)	QoG (ICRG)	CPI (TI)	Income (WDI)
Segregation	-1.40*	-1.14	-0.47	-1.24*	0.50	-0.14	-0.25	-2.01*
	(0.78)	(0.69)	(0.32)	(0.67)	(0.67)	(0.17)	(0.16)	(1.02)
R^2	0.38	0.43	0.35	0.43	0.50	0.46	0.40	0.49
Alignment	0.44**	0.43***	0.16**	0.40**	0.37**	0.08**	0.11**	0.48***
	(0.19)	(0.16)	(0.07)	(0.16)	(0.18)	(0.03)	(0.04)	(0.17)
Benchmark Seg.	-1.30	-1.12	-0.44	-1.27	0.83	-0.09	-0.22	-2.41**
	(0.91)	(0.80)	(0.36)	(0.77)	(0.77)	(0.20)	(0.18)	(1.12)
R^2	0.41	0.46	0.37	0.46	0.52	0.48	0.44	0.52
Alignment	0.39**	0.38**	0.14*	0.38**	0.36**	0.07**	0.10**	0.44**
	(0.20)	(0.17)	(0.08)	(0.15)	(0.17)	(0.03)	(0.04)	(0.20)
Fractionalization	-0.60	-0.44	-0.23	-0.59*	0.29	-0.04	-0.11	-1.06**
	(0.38)	(0.34)	(0.15)	(0.33)	(0.34)	(0.08)	(0.08)	(0.44)
Dispersion	-0.70	-0.85	-0.17	-0.38	-0.04	-0.13	-0.05	-0.59
	(0.97)	(0.79)	(0.42)	(0.78)	(0.85)	(0.19)	(0.21)	(1.23)
R^2	0.42	0.47	0.38	0.46	0.52	0.49	0.44	0.53
Continent FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Countries	157	157	157	157	157	131	154	152

Notes: Each column presents three OLS regressions with continent fixed effects. Dependent variables are control of corruption, government effectiveness, political stability, regulatory quality, and voice and accountability by the World Bank Governance Indicators in columns (1)–(5); quality of government by ICRG in column (6); the corruption perception index by Transparency International in column (7), and the log of real GDP per capita from the World Development Indicators in column (8). All dependent variables refer to 2010. Online Appendix E contains more information on the dependent variables. Segregation, alignment, benchmark segregation, fractionalization and dispersion are explained in Sections 2 and 3.1. Robust standard errors. ***, **, * indicate p-values below 0.01, 0.05 and 0.1, respectively.

Table G.5: Alternative computations of our indices of ethnic geography

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent var.	Rule of law (WBGI)			Income (PWT)			Trust (WVS)		
Segregation	-2.26***	-2.18***	-1.05	-1.97**	-2.00**	-0.81	0.05	0.02	0.49**
	(0.62)	(0.66)	(1.00)	(0.91)	(0.91)	(1.06)	(0.16)	(0.19)	(0.19)
R^2	0.40	0.39	0.36	0.52	0.52	0.51	0.25	0.25	0.31
Alignment	0.51***	0.72	0.55***	0.49**	0.76**	0.53***	0.10***	0.13*	0.10***
	(0.18)	(0.49)	(0.17)	(0.19)	(0.38)	(0.16)	(0.04)	(0.08)	(0.03)
Fract	-0.91***	-0.87**	-0.92***	-0.93**	-0.82*	-0.92**	0.14*	0.07	-0.00
	(0.30)	(0.36)	(0.35)	(0.39)	(0.42)	(0.43)	(0.07)	(0.09)	(0.09)
Dispersion	-0.86	-1.63	0.59	-1.00	-2.39	0.48	-0.16	-0.26	0.29***
	(0.84)	(1.49)	(0.58)	(1.19)	(1.72)	(0.54)	(0.18)	(0.31)	(0.07)
R^2	0.46	0.44	0.45	0.57	0.56	0.56	0.39	0.34	0.48
Continent FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Alternative ethnic distances	Yes	No	No	Yes	No	No	Yes	No	No
Non-linear spatial distances	No	Yes	No	No	Yes	No	No	Yes	No
Absolute spatial distances	No	No	Yes	No	No	Yes	No	No	Yes
Countries	157	157	157	148	148	148	77	77	77

Notes: Each column presents two OLS regressions with continent fixed effects. Dependent variables are the rule of law in 2010 by the World Bank Governance Indicators in columns (1)–(3), the log of expenditure-side real GDP per capita in 2010 from the Penn World Tables 9.0 in columns (4)–(6), and generalized trust from the World Value Survey in the 1981-2008 time period (Ashraf and Galor 2013) in columns (7)–(9). Appendix E contains more information on these variables. Segragation, alignment, fractionalization and dispersion are explained in Sections 2 and 3.1. However, we compute these indices slightly differently than reported in Section 3.1. We use ethnolinguistic distances calculated using the formula in Fearon (2003) in columns (1), (4) and (7); spatial distances as the square root of the relative geodesic distance in columns (2), (5) and (8); and absolute spatial distances in columns (3), (6) and (9). Robust standard errors. ***, **, * indicate p-values below 0.01, 0.05 and 0.1, respectively.

Table G.6: Allowing for non-linear effects of fractionalization and dispersion

	(1)	(2)	(3)
Dependent var.	Rule of law (WBGI)	Income (PWT)	Trust (WVS)
Alignment	0.47**	0.40**	0.10***
	(0.19)	(0.19)	(0.03)
Fractionalization	-2.89	-2.75	-0.15
	(1.96)	(2.04)	(0.48)
Fractionalization^2	0.53	-0.53	0.20
	(1.65)	(1.64)	(0.48)
Dispersion	-1.65	-11.10*	-0.03
	(4.39)	(5.64)	(0.95)
Dispersion^2	-0.30	11.68	-0.29
	(5.59)	(7.36)	(1.31)
Fractionalization	4.58	5.87	0.41
$\times$ Dispersion	(4.01)	(4.24)	(1.10)
Continent FE	Yes	Yes	Yes
R^2	0.45	0.58	0.40
Countries	157	148	77

Notes: OLS regressions with continent fixed effect. Dependent variables are the rule of law in 2010 by the World Bank Governance Indicators in column (1), expenditure-side real GDP per capita in 2010 from the Penn World Tables 9.0 in column (2), and generalized trust from the World Value Survey in the 1981-2008 time period (Ashraf and Galor 2013) in column (3). Online Appendix E contains more information on these variables. Alignment, fractionalization and dispersion are explained in Sections 2 and 3.1. The addition of square and interaction terms of fractionalization and dispersion allows showing that the coefficient on alignment is not driven by some non-linearity in the effects of fractionalization or dispersion. Robust standard errors. ***, **, * indicate p-values below 0.01, 0.05 and 0.1, respectively.

Table G.7: Weight least squares (WLS)

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent var.	Rule of law (WBGI)		Income (PWT)		Trust (WVS)	
Segregation	-1.40*	-1.37*	-1.08	-1.06	0.16	0.15
	(0.71)	(0.73)	(0.99)	(0.98)	(0.16)	(0.17)
R^2	0.43	0.43	0.55	0.55	0.36	0.37
Alignment	0.53***	0.49**	0.50***	0.46***	0.10***	0.09***
	(0.18)	(0.19)	(0.16)	(0.15)	(0.03)	(0.03)
Benchmark Seg.	-1.33	-1.28	-1.37	-1.36	0.34*	0.35*
	(0.82)	(0.84)	(1.12)	(1.10)	(0.19)	(0.19)
R^2	0.47	0.47	0.57	0.57	0.44	0.45
Alignment	0.51***	0.48***	0.45**	0.42**	0.09***	0.09***
	(0.17)	(0.18)	(0.18)	(0.17)	(0.03)	(0.03)
Fractionalization	-0.60	-0.59	-0.62	-0.63	0.15*	0.15*
	(0.36)	(0.37)	(0.44)	(0.44)	(0.08)	(0.09)
Dispersion	-0.40	-0.22	-0.77	-0.57	-0.08	-0.07
	(0.90)	(0.92)	(1.15)	(1.11)	(0.21)	(0.21)
R^2	0.47	0.47	0.58	0.58	0.45	0.46
Continent FE	Yes	Yes	Yes	Yes	Yes	Yes
Weights	Pop.	Area	Pop.	Area	Pop.	Area
Countries	157	157	148	148	77	77

Notes: Each column presents three WLS regressions with continent fixed effects. Weights are the log of population size in odd columns and the log of surface area in even columns, both from the World Development Indicators. Dependent variables are the rule of law in 2010 by the World Bank Governance Indicators in columns (1) and (2), expenditure-side real GDP per capita in 2010 from the Penn World Tables 9.0 in columns (3) and (4), and generalized trust from the World Value Survey in the 1981-2008 time period (Ashraf and Galor 2013) in columns (5) and (6). Online Appendix E contains more information on dependent and control variables. Segregation, alignment, benchmark segregation, fractionalization and dispersion are explained in Sections 2 and 3.1. Robust standard errors. ***, **, * indicate p-values below 0.01, 0.05 and 0.1, respectively.

Table G.8: Poisson pseudo-maximum likelihood (PPML)

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent var.	QoG (ICRG)		Income (PWT)		Trust (WVS)	
Alignment	0.20	0.31***	0.09***	0.14***	0.48***	0.54***
	(0.12)	(0.10)	(0.04)	(0.03)	(0.18)	(0.16)
Benchmark Seg.	-0.08***	-0.07***	-0.03***	-0.03***	-0.04	-0.02
	(0.02)	(0.02)	(0.01)	(0.01)	(0.04)	(0.03)
R^2	0.22	0.18	0.24	0.19	0.15	0.15
Alignment	0.23**	0.21**	0.10***	0.09***	0.49**	0.40**
	(0.11)	(0.09)	(0.04)	(0.03)	(0.19)	(0.16)
Fractionalization	-0.03	-0.02	-0.01*	-0.01*	0.02	0.01
	(0.02)	(0.02)	(0.01)	(0.00)	(0.04)	(0.02)
Dispersion	0.02	-0.09	-0.01	-0.05	-0.07	-0.14
	(0.13)	(0.12)	(0.04)	(0.04)	(0.22)	(0.22)
R^2	0.48	0.46	0.58	0.56	0.35	0.37
Continental dummies	Yes	Yes	Yes	Yes	Yes	Yes
Windsorizing F & BS	No	Yes	No	Yes	No	Yes
Countries	119	131	134	148	70	77

Notes: PPML regressions. Dependent variables are the quality of government by ICRG in columns (1) and (2), expenditure-side real GDP per capita in 2010 from the Penn World Tables 9.0 in columns (3) and (4), and generalized trust from the World Value Survey in the 1981-2008 time period (Ashraf and Galor 2013) in columns (5) and (6). Online Appendix E contains more information on these variables. We use the quality of government by ICRG rather than the rule of law in 2010 by the World Bank Governance Indicators as in most other tables, because PPML requires non-negative dependent variables. This change of the dependent variable leads to a drop in the sample size. All regressions include continental dummy variables. Alignment, fractionalization and dispersion all enter in logs. We thus lose all countries in which fractionalization is zero in odd columns. We add a small constant (0.001) to fractionalization and benchmark segregation before taking logs in even columns, which allows keeping these countries in the sample. Alignment, fractionalization and dispersion are explained in Sections 2 and 3.1. ***, **, * indicate p-values below 0.01, 0.05 and 0.1, respectively.

H. Cross-country regressions including alternative indices

Table H.1: Controlling for the standard fractionalization index by Alesina et al. (2003)

	(1)	(2)	(3)
Dependent variable	Rule of law (WBGI)	Income (PWT)	Trust (WVS)
Segregation (HVV)	-2.22***	-1.00	-0.00
	(0.82)	(1.25)	(0.25)
Standard fractionalization (ADEKW)	-0.04	-0.72	-0.03
	(0.34)	(0.45)	(0.08)
R^2	0.40	0.54	0.23
Alignment (HVV)	0.48***	0.43**	0.10***
	(0.18)	(0.18)	(0.04)
Generalized fractionalization (HVV)	-0.97**	-0.71	0.11
	(0.43)	(0.52)	(0.13)
Dispersion (HVV)	-0.80	-0.62	-0.15
	(0.88)	(1.23)	(0.20)
Standard fractionalization (ADEKW)	0.12	-0.46	-0.02
	(0.34)	(0.44)	(0.08)
R^2	0.45	0.57	0.37
Continent FE	Yes	Yes	Yes
Countries	153	145	75

Notes: Each column presents two OLS regressions with continent fixed effects. Dependent variables are the rule of law in 2010 by the World Bank Governance Indicators in column (1), expenditure-side real GDP per capita in 2010 from the Penn World Tables 9.0 in column (2), and generalized trust from the World Value Survey in the 1981-2008 time period (Ashraf and Galor 2013) in column (3). Segregation (HVV), alignment (HVV), generalized fractionalization (HVV) and dispersion (HVV) are our indices explained in Sections 2 and 3.1. Standard fractionalization (ADEKW) is the index of linguistic fractionalization based on categorical data as computed by Alesina et al. (2003). Robust standard errors. ***, **, * indicate p-values below 0.01, 0.05 and 0.1, respectively.

Table H.2: Controlling for the indices of standard fractionalization and a-spatial segregation by Alesina and Zhuravskaya (2011)

	(1)	(2)	(3)
Dependent variable	Rule of law (WBGI)	Income (PWT)	Trust (WVS)
Segregation (HVV)	0.06	0.63	0.13
	(1.04)	(1.37)	(0.34)
Standard fractionalization (AZ)	-0.29	-0.41	0.09
	(0.43)	(0.50)	(0.13)
Segregation (AZ)	-1.11	0.05	-0.26*
	(0.71)	(0.74)	(0.14)
R^2	0.44	0.64	0.31
Alignment (HVV)	0.68***	0.63***	0.10**
	(0.20)	(0.18)	(0.04)
Generalized fractionalization (HVV)	0.05	0.11	0.24
	(0.49)	(0.66)	(0.17)
Dispersion (HVV)	0.48	1.11	-0.13
	(1.49)	(0.97)	(0.27)
Standard fractionalization (AZ)	-0.11	-0.21	0.06
	(0.43)	(0.50)	(0.15)
Segregation (AZ)	-1.16	0.08	-0.27
	(0.72)	(0.73)	(0.18)
R^2	0.52	0.68	0.46
Continent FE	Yes	Yes	Yes
Countries	90	89	54

Notes: Each column presents two OLS regressions with continent fixed effects. Dependent variables are the rule of law in 2010 by the World Bank Governance Indicators in column (1), expenditure-side real GDP per capita in 2010 from the Penn World Tables 9.0 in column (2), and generalized trust from the World Value Survey in the 1981-2008 time period (Ashraf and Galor 2013) in column (3). Segregation (HVV), alignment (HVV), generalized fractionalization (HVV) and dispersion (HVV) are our indices explained in Sections 2 and 3.1. Standard fractionalization (AZ) is the index of linguistic fractionalization based on categorical data as computed by Alesina and Zhuravskaya (2011). Segregation (AZ) is the a-spatial segregation index used by Alesina and Zhuravskaya (2011), which is based on the population shares of different language groups in different subnational units rather than ethnolingustic and spatial distances. Robust standard errors. ***, **, * indicate p-values below 0.01, 0.05 and 0.1, respectively.

Table H.3: Controlling for the indices of standard fractionalization, generalized fractionalization and polarization by Esteban et al. (2012)

	(1)	(2)	(3)
Dependent variable	Rule of law (WBGI)	Income (PWT)	Trust (WVS)
Segregation (HVV)	-1.23	-2.51*	0.22
	(1.06)	(1.45)	(0.27)
Standard fractionalization (EMR)	-0.43	-0.51	-0.26***
	(0.41)	(0.48)	(0.08)
Generalized fractionalization (EMR)	0.15	1.50	0.61*
	(1.17)	(1.58)	(0.31)
Polarization (EMR)	-0.81	-1.20	-1.69**
	(3.01)	(3.81)	(0.70)
R^2	0.44	0.61	0.36
Alignment (HVV)	0.60***	0.46**	0.10***
	(0.18)	(0.18)	(0.04)
Generalized fractionalization (HVV)	-0.39	-1.17*	0.20
	(0.50)	(0.64)	(0.13)
Dispersion (HVV)	0.08	-1.26	-0.07
	(1.09)	(0.99)	(0.20)
Standard fractionalization (EMR)	-0.29	-0.29	-0.21***
	(0.38)	(0.46)	(0.07)
Generalized fractionalization (EMR)	-0.44	1.05	0.45
	(1.13)	(1.54)	(0.29)
Polarization (EMR)	0.91	0.36	-1.32*
	(2.91)	(3.69)	(0.67)
R^2	0.49	0.64	0.47
Continent FE	Yes	Yes	Yes
Countries	132	129	75

Notes: Each column presents two OLS regressions with continent fixed effects. Dependent variables are the rule of law in 2010 by the World Bank Governance Indicators in column (1), expenditure-side real GDP per capita in 2010 from the Penn World Tables 9.0 in column (2), and generalized trust from the World Value Survey in the 1981-2008 time period (Ashraf and Galor 2013) in column (3). Segregation (HVV), alignment (HVV), generalized fractionalization (HVV) and dispersion (HVV) are our indices explained in Sections 2 and 3.1. Standard fractionalization (EMR) is the index of ethnic fractionalization based on categorical data as computed by Esteban et al. (2012). Generalized fractionalization (EMR) is their Greenberg-Gini index, which is based on ethnic and spatial distances. Polarization (EMR) is the polarization index by Duclos et al. (2004) as computed by Esteban et al. (2012). Robust standard errors. ***, **, * indicate p-values below 0.01, 0.05 and 0.1, respectively.